Kwaliteit van het onderwijs gemeten : studies naar de betrouwbaarheid, validiteit en bruikbaarheid van studentoordelen

KWALITEIT VAN HET ONDERWIJS GEMETEN.

Studies naar de betrouwbaarheid, validiteit en bruikbaarheid van studentoordelen.

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de Rector Magnificus, Prof. Dr. F.I.M. Bonke,
volgens het besluit van het College van Dekanen,
in het openbaar te verdedigen op vrijdag,
9 september 1988 om 16.00 uur

door

Wilhelmus Hubertus Gijselaers

geboren te Heerlen in 1959

promotores:	Prof.Dr. H.G. Schmidt Prof.Dr. W.H.F.W. Wijnen
beoordelingscommissie:	Prof.Dr. J.J.C.B. Bremer Prof.Dr. B.P.M. Creemers Prof.Dr. H.F.M. Crombag Prof.Dr. D.W. Vaags Prof.Dr. G.J. van der Vusse

Voor Marjan

Omslag : Bert Kerkhofs

Tekstverwerking : Désirée Bronckers

Druk : Drukkerij Alberts/Druko Gulpen

ISBN nr. : 90-9002397-6

VOORWOORD.

In dit proefschrift wordt verslag gedaan van een aantal studies naar de betrouwbaarheid, validiteit en bruikbaarheid van studentoordelen. Het onderzoek waarover gerapporteerd wordt, werd verricht in het kader van het project 'Programma-evaluatie' dat uitgevoerd wordt door de vakgroep Onderwijsresearch en Onderwijsontwikkeling van de Rijksuniversiteit Limburg.
In de loop der jaren waren een aantal personen nauw betrokken bij dit project. Bij deze zou ik hen willen bedanken voor de raad en daad waarmee zij het project ondersteunden. In de eerste plaats betreft dat mijn collega's in de vakgroep Onderwijsontwikkeling en Onderwijsresearch. Dat geldt in het bijzonder voor Els Boshuizen, Willem de Grave, Jos Moust, Henk Schmidt, Hetty Snellen en Betsy Stalenhoef. Zij leverden als lid van de projectgroep een belangrijke bijdrage aan de dagelijkse voortgang en verdere ontwikkeling van het project.
Een bijzonder woord van dank verdient Bert Kerkhofs die als onderzoeksassistent de verantwoordelijkheid droeg voor de dataverzameling en dataverwerking. Sedert de start van het project analyseerde hij zonder morren vele malen de evaluatiegegevens. Bovendien bleek hij altijd bereid mee te denken over de wijze waarop de data het beste verwerkt konden worden.
Voorts wil ik Tjaart Imbos en Paul Jacobs danken. Tjaart Imbos verrichtte een aantal analyses voor studie 6 in hoofdstuk 4. Paul Jacobs werkte mee aan de opbouw en analyse van een aantal databestanden die in hoofdstuk 5 beschreven worden.
Buro Onderwijs van de medische faculteit verzorgde in de loop der jaren de administratieve ondersteuning van het project.
Tevens wil ik Geer Hoppenbrouwers en Désirée Bronckers bedanken voor de hulp die zij boden bij de totstandkoming van dit proefschrift. Geer Hoppenbrouwers las het conceptmanuscript en deed voorstellen voor de verbetering van de tekst en vormgeving van het typoscript. Désirée Bronckers verzorgde de tekstverwerking van het manuscript.
Voorts zou ik de leden van de beoordelingscommissie, bestaande uit Prof.Dr. J. Bremer, Prof.Dr. B. Creemers, Prof.Dr. H. Crombag, Prof.Dr. W. Vaags, en Prof.Dr. G. van der Vusse, willen danken voor hun kritische opmerkingen en suggesties.
Tenslotte dank ik de promoteren Prof.Dr. H. Schmidt en Prof.Dr. W. Wijnen voor de wijze waarop zij mij de afgelopen jaren begeleid hebben bij het schrijven van dit proefschrift.

Mheer, juli 1988.

Wim Gijselaers.

INLEIDING.

Universiteiten zijn de laatste jaren steeds meer aandacht gaan besteden aan de kwaliteit van het onderwijs. De vraag die zich daarbij onmiddellijk voordoet is hoe onderwijskwaliteit gemeten kan worden. Een andere vraag is op welke wijze onderwijskwaliteit verbeterd kan worden. In dit proefschrift wordt een onderzoek beschreven waarin aan beide aspecten aandacht geschonken wordt. Dit onderzoek vond plaats binnen de Faculteit der Geneeskunde van de Rijksuniversiteit Limburg. Binnen deze faculteit is door het project programma-evaluatie een evaluatie-aanpak ontwikkeld waarmee systematisch het onderwijsprogramma geëvalueerd wordt. Deze aanpak is met name gericht op het meten van het functioneren van het onderwijsprogramma.
Het project programma-evaluatie heeft als taak docenten en bestuurders van informatie te voorzien, teneinde de kwaliteit van het onderwijs te kunnen controleren en te verbeteren. Binnen de gehanteerde evaluatie-aanpak wordt voornamelijk gebruik gemaakt van vragenlijsten. Daarmee worden oordelen van studenten en docenten verzameld over de kwaliteit van het onderwijs. Twee factoren spelen een rol bij de realisatie van deze doelstelling: de beschikbaarheid van betrouwbare en valide gegevens en inzicht in het gebruik van die gegevens. In dit proefschrift wordt aan beide factoren aandacht besteed. Het proefschrift bevat in de eerste plaats een aantal empirische studies naar de vraag of de gehanteerde meetmethode betrouwbare en valide gegevens oplevert. In de tweede plaats, maar daarom niet minder belangrijk, worden een aantal studies beschreven naar de vraag in hoeverre de gegevens bruikbaar zijn voor de onderwijspraktijk.
Het proefschrift bestaat uit 5 hoofdstukken. De kern van het proefschrift wordt gevormd door de hoofdstukken 4 en 5. In deze hoofdstukken wordt het empirisch onderzoek naar de betrouwbaarheid, validiteit en bruikbaarheid van de evaluatiegegevens beschreven. In hoofdstuk 1 wordt aandacht besteed aan een aantal theoretische en praktische vraagstukken op het gebied van evaluatie in het hoger onderwijs. Hoofdstuk 2 bevat een beschrijving van het gehanteerde meetinstrument. Hoofdstuk 3 geeft een overzicht van meetproblemen die zich kunnen voordoen bij het gekozen instrument.

HOOFDSTUK 1. EVALUATIE IN HET HOGER ONDERWIJS.

1.1 Inleiding.

De laatste jaren valt in Nederland een toenemende belangstelling te constateren voor de kwaliteit van het product dat door het universitaire onderwijs geleverd wordt. Veranderingen in de wet- en regelgeving ten aanzien van het universitaire onderwijs en bezuinigingen op de overheidsuitgaven ten behoeve van dat onderwijs hebben ertoe geleid, dat universiteiten meer aandacht moeten besteden aan de bewaking en de verbetering van de kwaliteit van het onderwijs. Sedert de nota Posthumus (1968) heeft de overheid namelijk een groot aantal beslissingen genomen met als voornaamste doelstelling een verlaging van de kosten van het universitaire onderwijs, gepaard aan het constant houden en het zo mogelijk verhogen van de kwaliteit het onderwijs. De wet "Tweefasenstructuur" (1979), het nieuwe "Academisch Statuut" (1981), de "Wet op het Wetenschappelijk Onderwijs" (1981) en de beleidsnota's "Hoger Onderwijs: Autonomie en Kwaliteit", HOAK-nota, (1985) en "Hoger Onderwijs en Onderzoeksplan", HOOP-nota, (1987), beogen deze doelstellingen te bereiken door ingrijpende hervormingen in het universitaire onderwijs door te voeren zoals verkorting van inschrijvingsduur, reorganisatie van bestuursstructuren, herstructurering van de financiering van onderzoek en onderwijs (o.a. financiering op basis van output) en modularisering van het onderwijs (vouchersysteem).

In de HOOP-nota (1987) wordt de gedachte geformuleerd dat, conform de ideeën in de HOAK-nota (1985), instellingen zelfstandig afwegingen kunnen maken over al dan niet wenselijke ontwikkelingen in profiel, onderwijs en onderzoek. Een toegenomen zelfstandigheid kan echter pas gerealiseerd worden als de instellingen verantwoording zullen afleggen over de resultaten van hun aktiviteiten. Op basis van de strategie 'sturing op afstand' wil de overheid instellingen meer zelfstandigheid geven, mits deze achteraf verantwoording afleggen aan de geldgevers (overheid of derden) over de inzet van financiële en materiële middelen in relatie tot de gerealiseerde produktie. In tegenstelling tot de huidige besturingswijze, waarbij procedurele voorschriften de kwaliteit van het onderwijs vooraf moeten garanderen, wordt in de nabije toekomst een systeem ingevoerd waarbij de onderwijskwaliteit achteraf door de overheid gereguleerd wordt (HOOP, 1987).
Een dergelijke verandering in de regulering van het hoger onderwijs heeft voor universitaire instellingen ingrijpende gevolgen. Universiteiten moeten namelijk systemen voor interne kwaliteitsbewaking gaan opzetten die informatie verschaffen waarmee verantwoording afgelegd kan worden aan instellingsoverstijgende evaluatie- of visitatiecommissies. Deze commissies worden geacht een centrale rol te gaan spelen in het systeem van externe kwaliteitsbewaking. In deze commissies hebben externe deskundigen zitting ter beoordeling

van de kwaliteit van het onderwijs op basis van door de universiteiten te verschaffen informatie en door middel van ter plaatse te voeren gesprekken. In deze opzet zijn universiteiten zelf verantwoordelijk voor de uitvoering van deze vorm van externe kwaliteitsbewaking. Zij dragen via de VNSU (Vereniging van Nederlandse Samenwerkende Universiteiten) zorg voor de instelling van de commissies. Visitatiecommissies zullen hun oordeel vooral gaan baseren op de uitkomsten van zogenaamde "prestatie-indicatoren" (bijvoorbeeld cijfers over het aantal ingeschreven studenten, het numeriek rendement van de opleiding, de inzet van het wetenschappelijk personeel) en op informatie die door de faculteiten zelf wordt aangedragen. Hun taak komt sterk overeen met die van in de V.S. al langer bestaande accrediteringscommissies, zij het dan dat -op dit moment- geen formele bevoegdheid bestaat om faculteiten de wacht aan te zeggen, wanneer daarvoor redenen zouden zijn. De inspectie voor het hoger onderwijs speelt in dit geheel een aanvullende rol. Zij bewaakt of de eigen evaluatie van de instellingen door middel van periodieke visitatie, aan de eisen voldoet. De invoering van dit systeem van externe kwaliteitsbewaking heeft tot gevolg dat faculteiten genoodzaakt worden om, meer dan voorheen, aandacht te besteden aan de kwaliteit van hun onderwijs. Interne kwaliteitsbewaking is in academisch Nederland dan ook pas sinds kort in de belangstelling gekomen. In de jaren zestig en zeventig verschenen in de Nederlandstalige literatuur over dit onderwerp slechts enkele publikaties en dan nog uitsluitend in onderwijskundige vaktijdschriften (zie voor een overzicht Blom & Langerak, 1979). Pas de laatste jaren wordt over dit onderwerp meer gepubliceerd (bijvoorbeeld Bartelds, Joostens, Kluiter, 1983; Daniëls, 1985; Van Os, 1987).

De ontwikkelingen die zich nu in Nederland voordoen (inspectie op het hoger onderwijs, visitatiecommissies) waren in de Verenigde Staten reeds dertig jaar geleden aan de orde. Onvrede over het bestaande onderwijsniveau, die mede in de hand gewerkt werd door de schok die de lancering van de eerste Russische Spoetnik in 1957 teweegbracht, had tot gevolg dat men steeds meer aandacht ging schenken aan kwaliteitsbeoordeling en dat de zorg om uitzonderlijke prestaties toenam. Het resultaat van deze ontwikkeling is ondermeer zichtbaar in het feit dat thans bijna iedere universiteit in de V.S. een uitgebreid systeem voor interne kwaliteitsbewaking kent. De ervaringen daarmee zijn uitermate bruikbaar voor het ontwikkelen van evaluatiesystemen ten behoeve van Nederlandse universiteiten. We zullen in dit proefschrift daarom veelvuldig verwijzen naar de situatie in de Verenigde Staten.

Binnen de Rijksuniversiteit Limburg wordt, sedert haar oprichting in 1974, het onderwijs systematisch in alle faculteiten geëvalueerd. Daarmee vormt zij wellicht een uitzondering als het gaat om systematische evaluatie van universitaire onderwijsprogramma's. Dat heeft ook een duidelijke oorzaak. Deze universiteit heeft namelijk sinds 1974 op grote

schaal onderwijsvernieuwing nagestreefd, zowel wat betreft het onderwijsmanagement (het gebruiken van een matrixorganisatievorm), als wat betreft de gehanteerde didactische werkvormen (probleemgestuurd onderwijs) en de wijze van examinering (voortgangstoetsen, bloktoetsen). Daardoor ontstond vanaf het begin bij de betrokkenen behoefte aan informatie over het functioneren van de onderwijsprogramma's. Onzekerheid en twijfels over de vraag of deze vernieuwing wel tot de gewenste resultaten zou leiden, bracht de faculteiten ertoe hun onderwijs systematisch te evalueren. Het doel dat men daarbij voor ogen had, was tweeledig: 1) het bijstellen en verbeteren van programma's en 2) het afleggen van verantwoording aan derden (zusterfaculteiten, de overheid).
In dit proefschrift wordt in het bijzonder aandacht besteed aan een systeem voor programma-evaluatie dat aan de Rijksuniversiteit Limburg ontwikkeld is. Dit systeem (naar een ontwerp van H. G. Schmidt , 1981) is door de projectgroep Programma-evaluatie, binnen de vakgroep Onderwijsontwikkeling en Onderwijsresearch in eerste instantie ontwikkeld ten behoeve van de Faculteit der Geneeskunde. Het beoogt docenten, vak- en projectgroepen, facultaire bestuursorganen en andere bij het onderwijs betrokkenen, informatie te geven over het functioneren van het onderwijsprogramma teneinde de kwaliteit van het onderwijs te verbeteren. Programma-evaluatie kan binnen dit kader gezien worden als een instrument voor interne kwaliteitsbeheersing.
In dit proefschrift worden drie vragen met betrekking tot dit systeem behandeld: wat is de betrouwbaarheid, de validiteit en de bruikbaarheid van de informatie die binnen dat systeem verzameld wordt? Het proefschrift bestaat uit vijf hoofdstukken waarin theoretische en praktische vraagstukken behandeld worden die te maken hebben met het vergaren, het analyseren en het effectief gebruik maken van informatie over onderwijsprocessen.
In de volgende paragrafen wordt achtereenvolgens aandacht besteed aan de beoordelingscriteria die gehanteerd kunnen worden om de kwaliteit van onderwijsprogramma's vast te stellen, aan de meetinstrumenten die beschikbaar zijn om kwaliteit te meten, en aan de condities waaronder evaluatie een bijdrage kan leveren aan de verbetering van de kwaliteit van het onderwijs.

1.2 Opvattingen over het definiëren en meten van de kwaliteit van het universitaire onderwijs.

In dit proefschrift houden we ons, zoals gezegd, bezig met de vraag in hoeverre met behulp van geformaliseerde vormen van evaluatie, de kwaliteit van onderwijsprogramma's beoordeeld en verbeterd kan worden. De verzamelde informatie moet daarbij daadwerkelijk aanleiding geven tot verandering en verbetering. Wanneer men zich een dergelijke taak stelt, is uiteraard de eerste vraag, die zich onmiddellijk aandient, hoe men de kwaliteit van onderwijs kan definiëren. De tweede vraag is welke methoden of evaluatie-aanpakken bestaan om onderwijskwaliteit te meten. In deze paragraaf zal daarom achtereenvolgens aandacht besteed worden aan de keuze van beoordelingscriteria en aan het meten van de kwaliteit van het onderwijs.

1.2.1 Enkele opvattingen over het begrip "kwaliteit van het onderwijs".

In de literatuur worden verschillende invullingen gegeven aan het begrip onderwijskwaliteit. Giesbers (1986) stelt ondermeer dat het weinig zinvol is om van de kwaliteit van onderwijs te spreken. Het begrip heeft volgens hem een multidimensionaal en context-afhankelijk karakter. Invulling van het begrip verschilt naar gelang van de vraag of het betrekking heeft op doelstellingen, werkvormen, inhouden, of produkten van onderwijs. Afhankelijk van de organisatorische context waarin de evaluatie plaatsvindt, bijvoorbeeld op het niveau van de instelling, de faculteit, of de vakgroep, zal het begrip onderwijskwaliteit een andere invulling krijgen. Er bestaan met andere woorden in zijn visie -en wij sluiten ons daarbij aan- geen eenduidige criteria die in alle mogelijke situaties toepasbaar zijn, om de kwaliteit van het onderwijs te definieren. In de nota Kwaliteit van het wetenschappelijk onderwijs (1985) maakt de Academische Raad onderscheid tussen interne en externe kwaliteit van het onderwijs.
Externe kwaliteit heeft betrekking op de deugdelijkheid, doelmatigheid, doeltreffendheid en bruikbaarheid van onderwijs, gezien vanuit het perspectief van de gebruiker (studenten) en de afnemer (samenleving). Interne kwaliteit verwijst naar kenmerken van het onderwijsproces (voorbereiding, uitvoering en evaluatie), gezien vanuit het perspectief van de onderwijsgever.
Conrad en Blackburn (1985) trekken in een artikel dat handelt over externe kwaliteitsbeoordeling in de Verenigde Staten, vergelijkbare conclusies als Giesbers (1986). Zij komen eveneens tot de slotsom dat over het begrip "kwaliteit van het onderwijs" geen universele overeenstemming bestaat. Volgens deze auteurs worden meestal tamelijk globale criteria gebruikt, zoals: 'accountability', "efficiency", effectiviteit en "excellence". De term accountability kan vertaald worden als verantwoording tegenover de geldverstrekker.

Een onderwijsprogramma is 'accountable' als daarin een bepaald niveau nagestreefd wordt en door een voldoende aantal studenten wordt bereikt. Kwaliteit verwijst dus ondermeer naar de keuze van de doelstellingen en het tot op bepaalde hoogte bereiken van deze doelstellingen. "Execellence" heeft betrekking op de vraag hoeveel gerenommeerde docenten onderwijs geven in een programma. Efficiency en effectiviteit zijn economische begrippen die de doelmatigheid en doeltreffendheid van een opleiding betreffen. Onderwijs is effectief naarmate de vooropgezette doelstellingen bereikt worden. Het is efficiënt als die doelstellingen bereikt worden met een minimum aan kosten en inspanning.

1.2.2 Procedures om de kwaliteit van het onderwijs te meten.

De hierboven genoemde criteria zijn voorbeelden van globale referentiekaders waaraan de kwaliteit van een opleiding getoetst kan worden. Een probleem is echter dat deze kaders niet specifiek genoeg zijn om informatie te verschaffen die in de praktijk gebruikt kan worden om de kwaliteit van het onderwijs te verbeteren. Er zijn dan ook verschillende evaluatiebenaderingen ontwikkeld die deze criteria ieder op een andere manier operationaliseren, en die derhalve als instrument voor interne kwaliteitsbeheersing kunnen fungeren. Men kan grofweg vier evaluatiebenaderingen onderscheiden die veelvuldig binnen het universitair onderwijs toegepast worden: 1) doelstelling-georienteerd, 2) "characteristics of teaching effectiveness", 3) "models of teaching", en 4) "illuminative evaluation".

Doelstelling-georiënteerde evaluatie.

Doelstelling-georiënteerde evaluatie concentreert zich met name op het meten van de effectiviteit van onderwijsprogramma's, d.w.z. op de vraag in hoeverre de vooropgezette doelen worden bereikt. Tyler (1934) wordt als de grondlegger van deze benadering gezien. Hij ontwikkelde een evaluatiemodel waarin achtereenvolgens zes stappen afgehandeld moesten worden om een programma te evalueren: 1) het formuleren van doelstellingen, 2) het classificeren van doelstellingen, 3) de operationalisatie van doelstellingen in observeerbaar gedrag, 4) het specificeren in welke situaties dit gedrag aangetoond zou kunnen worden, 5) het ontwikkelen van meetinstrumenten, en 6) het verrichten van voor- en nametingen. In figuur 1.1 wordt geïllustreerd dat de doelstelling-georienteerde aanpak zich exclusief richt op de relatie tussen doelen en product.
Andere factoren die het onderwijsproces kunnen beïnvloeden, bijvoorbeeld de gehanteerde werkvorm en de aard van leerstof, worden buiten beschouwing gelaten.

Fig 1.1: Doelstelling-georienteerde benadering.

DOELEN ----------> PRODUCT

Bloom (1970) ontwikkelde deze aanpak verder, met name wat betreft het aspect van de ontwikkeling van meetinstrumenten. Hij definieerde evaluatie als: 'a process for determining the achievement of specific educational objectives'. Studieprestaties of toetsresultaten waren volgens Bloom (1970) goede indicatoren voor het meten van de effectiviteit van cursussen of programma's. De mate waarin de doelstellingen bereikt worden, kan binnen deze aanpak bepaald worden door prestaties op toetsen die als operationalisatie van de vooropgezette doelen zijn te beschouwen, te vergelijken met de doelen. Bloom (1959) ontwikkelde daartoe een taxonomie om onderwijsdoelstellingen te identificeren, te beschrijven, te classificeren en te meten. Deze aanpak heeft twee belangrijke voordelen. In de eerste plaats worden docenten gedwongen om precies rekenschap te geven welke doelen met het onderwijs nagestreefd worden. Blooms aanpak vereist dat in de beschrijving van een onderwijsprogramma eindtermen worden opgenomen. Daarnaast moeten de onderwijsdoelstellingen omschreven worden in termen van observeerbaar gedrag. Deze eisen vergemakkelijkten het meten van de effecten van onderwijsprogramma's aanzienlijk. De nadelen zijn echter ook evident: studieprestaties geven weliswaar inzicht in de zwakke plekken van een onderwijsprogramma, maar ze geven geen <u>verklaring</u> voor deze tekorten (De Corte, 1976). Bovendien vormen toetsuitkomsten eerder een maat voor studentactiviteiten, dan een maat voor de activiteiten van de docent. Om inzicht te krijgen in de wijze waarop studieprestaties tot stand komen, is het noodzakelijk dat men aandacht heeft voor wat er <u>tijdens</u> het onderwijs allemaal gebeurt.

In reactie op de doelstelling-georienteerde aanpak werden benaderingen ontwikkeld die een grotere nadruk legden op het beschrijven van het onderwijsproces. Immers om beslissingen te nemen over een onderwijsprogramma met als doel de onderwijskwaliteit te verbeteren, is het noodzakelijk dat men over informatie beschikt die de keuze tussen beschikbare alternatieven vergemakkelijkt. Evaluatie moet een bijdrage leveren in dit beslissingsproces door gegevens beschikbaar te stellen over de kwaliteit van het onderwijsproces en de wijze waarop het onderwijsproces het onderwijsproduct beinvloedt. Bijvoorbeeld door een antwoord te vinden op de volgende vragen: "Sluit het leermateriaal aan op de voorkennis van studenten?" "Worden practica en lezingen op meest geschikte plaats in een cursus gegeven?" "Sluiten de toetsen aan op de behandelde leerstof?" "Is er voldoende tijd voor zelfstudie beschikbaar?" "Wordt de behandelde leerstof als relevant ervaren door de studenten?" De hieronder te bespreken benaderingen

kunnen als conceptualiseringen van deze opvatting gezien worden.

"Characteristics of teacher's effectiveness" benadering.

In de doelstelling-georienteerde benadering lag de nadruk op de relatie tussen de vooropgezette doelen en de bereikte doelen. De effectiviteit van het onderwijs werd bepaald door de mate van overeenkomst tussen het geplande en daadwerkelijk bereikte. In de hieronder beschreven aanpak ligt de nadruk op het onderwijsleerproces en met name op de rol van de docent daarin. Men veronderstelt, dat datgene wat de docent in het onderwijs doet, weerspiegeld worden in de resultaten van de student. Kenmerken van het onderwijsproces zijn dus gerelateerd aan het onderwijsproduct, bijvoorbeeld van studieprestaties, of van de motivatie van studenten. Met andere woorden, de kwaliteit van het onderwijs kan bepaald worden door het onderwijsproces, in casu het gedrag van de docent, te observeren. Vragen die zich daarbij natuurlijk onmiddellijk voordoen zijn: Welke criteria moeten worden gehanteerd om het proces te beoordelen? Wat is een goede docent? Wat is goede leerstof? Welke werkvormen zijn in een gegeven situatie vereist? Een vraag die hierop aansluit is welke docenten positievere effecten hebben op de studieresultaten van studenten dan andere. In figuur 1.2 wordt getoond dat deze aanpak zich vooral richt op de relatie tussen instructie en product en in het bijzonder op de rol van de docent daarin. Er wordt, in tegenstelling tot de eerder beschreven doelstelling-georiënteerde benadering, geen aandacht besteed aan de relatie tussen geplande doelen en product.

Figuur 1.2 "Characteristics of teacher effectiveness" benadering.

Een beperking van deze benadering is dat deze zich vooral concentreert op de docent en zijn handelen, en weinig aandacht besteedt aan de andere elementen van het onderwijsproces. In onderzoek richt men zich dus vooral op indicatoren waarmee goede van slechte docenten onderscheiden kunnen worden. Dit probleem wordt meestal op een inductieve manier aangepakt (Braskamp, Brandenburg & Ory, 1984). Men inventariseert op grote schaal opinies van studenten, docenten en bestuurders, over wat zij onder een goede docent, of een goede cursus verstaan. Aan studenten wordt bijvoorbeeld gevraagd een beschrijving te geven van de ideale docent. Aan docenten vraagt men op welke punten zij zouden letten als men de wijze waarop een collega lesgeeft, zouden moeten beoordelen. De

resultaten van deze inventarisaties hebben hun weerslag gevonden in de constructie van beoordelingsvragenlijsten, waarin vragen gesteld werden over de wijze van collegegeven door de docent (structureert hij/zij de leerstof, sluiten de lessen aan op de voorkennis van studenten, toont hij/zij voldoende interesse voor eventuele studieproblemen van studenten, etc.).
Dit soort onderzoek is met name in de Verenigde Staten veelvuldig uitgevoerd. Het heeft een aantal instrumenten opgeleverd waarmee de kwaliteit van docenten beoordeeld kan worden. Deze lijsten bevatten items die, een operationalisatie vormen van kenmerken van goede docenten. In tabel 1.1 zijn als voorbeeld een aantal items uit zo'n lijst overgenomen.

Tabel 1.1 Items ontleend aan het 'Student Instructional Report'. Uit Braskamp, Brandenburg & Ory (1984).

	NA	SA	A	D	SD
1. The instructor's objectives for the course have been made clear.....	0	0	0	0	0
2. The instructor used class time well.....	0	0	0	0	0
3. The instructor seemed to know when students didn't seem to understand the material.....	0	0	0	0	0
4. The instructor made helpful comments on papers or exams.....	0	0	0	0	0

NA = Not Applicable, A = Agree, SD = Strongly Disagree.
SA = Strongly Agree, D = Disagree,

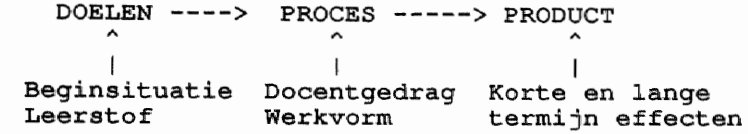
Hoewel deze benadering ontegenzeggelijk inzicht heeft verschaft in het functioneren van docenten binnen hun onderwijssetting, vertoont ze ook zwakke kanten. Onderzoek naar het verband tussen het gedrag van docenten en studieresultaten van studenten heeft namelijk herhaaldelijk aangetoond dat studieprestaties slechts gedeeltelijk bepaald worden door wat de docent doet gedurende het onderwijs (Cohen, 1981; Marsh, 1984). Allerlei andere factoren die de studieprestaties van studenten kunnen beïnvloeden (de invloed van andere vakken op de zelfstudie van studenten, de motivatie en voorkennis van studenten, de gepercipieerde relevantie van de cursus, de zwaarte van de cursus, etc.), worden buiten beschouwing gelaten. Men beperkt zich immers tot het observeren en meten van docentengedrag. Aktiviteiten van studenten naar aanleiding van docentaktiviteiten worden genegeerd.

"Models of teaching"-benadering.

In de "models of teaching" benadering, die men in tegenstelling tot die van de inductieve "teachers" effectiveness deductief zou kunnen noemen, richt men zich eveneens op het

onderwijsproces en zijn effecten. Men formuleert hier echter criteria op grond van theorieën over het onderwijsleerproces en niet op grond van opinies van bij het onderwijs betrokkenen. Tevens worden meer factoren dan alleen het gedrag van de docent in de beschouwing betrokken. Factoren die de gehele procesgang van het onderwijs (voorbereiding, uitvoering en evaluatie) beïnvloeden, worden geobserveerd en onderling met elkaar in verband gebracht. Voorbeelden van dergelijke factoren zijn: de aard van de beginsituatie (motivatie en voorkennis van de studenten, leerstof, docentkenmerken), kenmerken van het onderwijsproces (werkvorm, aktiviteiten docent en studenten) en de leerresultaten van de studenten. In figuur 1.3 is aangegeven op welke aspecten van het onderwijs deze aanpak gericht is.

Figuur 1.3 "Models of teaching" benadering.



De "models of teaching"-aanpak, heeft een sterk modelmatig karakter. De identificatie van variabelen die een rol spelen in het onderwijsleerproces en hun onderlinge causale relaties, krijgt erg veel nadruk (Carroll, 1963; Cooley & Leinhardt, 1980). In hoofdstuk 2 zullen we uitgebreider ingaan op enkele van deze modellen.

"Illuminative evaluation"-benadering.

Parlett en Hamilton (1977) beschrijven een evaluatie-aanpak die sterk afwijkt van de hierboven beschreven benaderingen. Zij constateren dat aan de klassieke evaluatiebenaderingen, waaronder de hierboven genoemde, drie tekortkomingen kleven. In de eerste plaats is er een te grote scheiding tussen de werkelijkheid en de assumpties die ten aanzien van die werkelijkheid gehanteerd worden. Het a priori opleggen aan de werkelijkheid van een aantal restricties leidt tot al te grote simplificaties, waardoor evaluatiegegevens betekenisloos kunnen worden. De auteurs geven als voorbeeld dat doelstellinggeoriënteerde evaluaties een te grote nadruk leggen op de specificatie van geplande doelen en de operationalisatie daarvan in studietoetsen. Sommige doelen zijn volgens hen te complex om zodanig omschreven te worden dat daaruit studietoetsen afgeleid kunnen worden. In de tweede plaats wordt teveel aandacht besteed aan de toepassing van geformaliseerde evaluatiedesigns. Het onderzoeksdesign krijgt teveel prioriteit waardoor kunstmatige, onnatuurlijke situaties in het onderwijs ontstaan. Docenten mogen bijvoorbeeld niet

afwijken van de voorgeschreven leerstof omdat dan het design aangetast wordt. In de derde plaats wordt slechts een beperkt aantal factoren geobserveerd die van invloed zijn op het onderwijs. Derhalve is de kans groot dat atypische resultaten en onverwachte effecten van het onderwijs niet waargenomen worden.
Om aan bovengenoemde bezwaren tegemoet te komen ontwikkelden Parlett en Hamilton (1977) een aanpak die bij uitstek geschikt is voor de evaluatie van bepaalde soorten onderwijsprogramma's: 1) programma's die complexe doelstellingen nastreven die moeilijk operationaliseerbaar zijn in observeerbaar gedrag, 2) programma's die afwijkende doelstellingen nastreven, of afwijkende werkvormen hanteren en waarvoor geen standaardinstrumenten hanteerbaar zijn, en 3) programma's waarin geen standaard evaluatiedesigns toegepast kunnen worden bij gebrek aan tijd, geld of bepaalde gegevens.
Bij illuminatieve evaluatie ligt het accent van de aktiviteiten van de evaluator op het verhelderen van de manier waarop het onderwijs functioneert en niet op het toetsen van leerresultaten of docentengedrag aan bepaalde vooraf gespecificeerde criteria. Normaal geeft een vakgroep of faculteit opdracht tot evaluatie van het onderwijs. De directe behoefte aan informatie over het functioneren van het onderwijs ligt echter bij degene die het onderwijs voorbereidt en uitvoert: de individuele docent. Veelal is de aangedragen informatie echter voornamelijk toegesneden op de wensen van bestuurders. Parlett en Hamilton leggen juist de nadruk op het verhelderen en beschrijven van het programma in uitvoering, zodat de verantwoordelijke docent de informatie kan gebruiken om het programma te verbeteren.
Het leermilieu en het onderwijssysteem fungeren als sleutelbegrippen bij illuminatieve evaluatie. Onder het leermilieu wordt de sociaal-psychologische omgeving verstaan waarin studenten en docenten met elkaar samenwerken. Aan het onderwijssysteem liggen bepaalde pedagogische assumpties ten grondslag die geformaliseerd zijn in beleidsnota's, cursusboeken, etc.. Het onderwijssysteem en het leermilieu zijn volgens de eerder genoemde auteurs van invloed op de uitkomsten van onderwijs. De evaluator moet derhalve over beide zaken informatie verzamelen. Hun onderzoeksinstrumentarium omvat observatie, interviews, vragenlijsten en documenten.
De illuminatieve evaluatie omvat drie fasen: observatie of probleemverkenning, onderzoeksuitvoering en het vinden van verklaringen. Deze fasen vormen een empirische cyclus waarin de onderzoeksvraagstelling steeds weer bijgesteld kan worden.
De evaluator staat tijdens het onderzoek in voortdurende wisselwerking met het onderwijs en kan daardoor, in tegenstelling tot de vorige benaderingen, niet als een onafhankelijke externe observator beschouwd worden.

Iedere onderwijsinstelling wordt vroeg of laat geconfronteerd met de vraag welke benadering het meest geschikt lijkt om haar onderwijsprogramma te evalueren, zo ook de medische

faculteit van de Rijksuniversiteit Limburg. In paragraaf 1.1 werd opgemerkt dat programma-evaluatie binnen de medische faculteit een instrument is om de kwaliteit van het onderwijs te meten en te verbeteren. Bij de keuze van een evaluatiebenadering kan men zich derhalve laten leiden door twee principes. Het eerste principe is dat men zich eerst afvraagt welke benadering het meest geschikt is om onderwijskwaliteit te meten, om vervolgens de vraag te stellen of deze geschikt is om de kwaliteit te verbeteren. De keuze voor een evaluatiebenadering wordt dan grotendeels afhankelijk gesteld van de vraag welke benadering kwaliteit het beste lijkt te meten. Het tweede principe start bij de vraag welke benadering het meest geschikt is de onderwijskwaliteit te verbeteren en heeft als eindpunt de meetvraag.
Darling-Hammond, Wise en Pease (1983) doen een poging om een model te ontwikkelen waarin beide principes gecombineerd worden om tot een geschikte keuze te komen. In dat model worden de verschillende evaluatiebenaderingen gerangschikt naar twee dimensies: 1) de uitgangspunten die gehanteerd worden bij het meten van de kwaliteit van het onderwijs, en 2) het soort informatie dat verstrekt wordt aan degenen die beslissingen nemen over het onderwijs.
Met betrekking tot de eerste dimensie worden evaluatiebenaderingen onderverdeeld in meer modelmatige en meer holistische benaderingen. Volgens deze auteurs liggen aan de eerder geschetste evaluatiebenaderingen namelijk twee tegengestelde opvattingen over onderwijs ten grondslag. De eerste opvatting is dat onderwijs volgens een planmatig en rationeel proces verloopt. Volgens dat idee kan onderwijs bestuurd worden door achtereenvolgens doelen te kiezen, daarna werkvormen voor te schrijven waarmee deze doelen bereikbaar zijn, en tenslotte te evalueren of deze doelen daadwerkelijk bereikt zijn. Onderwijs wordt binnen dit kader opgevat als een technologie die rationeel bestuurbaar is. Dergelijke opvattingen liggen mede ten grondslag aan de eerste drie hierboven genoemde evaluatiebenaderingen. Vanuit deze benaderingen is evaluatie gericht op het meten van variabelen die de kwaliteit van het onderwijs determineren. Vooraf kan gespecificeerd worden welke waarden deze variabelen moeten aannemen om te spreken van relatief goed of slecht onderwijs. De tweede opvatting, die onder andere verwoord wordt door Parlett en Hamilton (1977), behelst dat onderwijs niet altijd planmatig verloopt en dat ook andere, niet-meetbare of niet-verwachte factoren het onderwijs kunnen beïnvloeden. Het is volgens deze opvatting dus niet altijd mogelijk om restricties aan de werkelijkheid op te leggen, omdat de werkelijkheid zich afwijkend en onvoorspelbaar kan gedragen.
Met betrekking tot de tweede dimensie maken zij een indeling in evaluatiebenaderingen die meer gericht zijn op het geven van informatie aan de direct bij het onderwijs betrokkenen (studenten en docenten), en in evaluatiebenaderingen die meer gericht zijn op het geven van informatie aan degenen die verantwoordelijk zijn voor het onderwijsbeleid (facultaire

bestuurders). De illuminatieve evaluatiebenadering van Parlett en Hamilton (1977) is bijvoorbeeld vooral gericht op het concreet beschrijven van het programma-in-uitvoering. Daarmee beoogt men docenten informatie te verschaffen die direct bruikbaar is voor het verbeteren van één bepaalde cursus. De andere hierboven genoemde benaderingen zijn daar minder direct op gericht. De daarmee beoogde informatie is vooral voor bestuurders van belang omdat zij cursusoverstijgend is en daardoor minder betekenis heeft voor individuele docenten

Aan het model van Darling-Hammond, Wise en Pease (1983) ligt de gedachte ten grondslag, dat docenten en bestuurders bepaalde ideeën of concepties hebben over het onderwijsproces die zich vervolgens uitkristalliseren in de vorm en inhoud van het onderwijsprogramma (bijvoorbeeld de gehanteerde onderwijsfilosofie) en in de structuur van de onderwijsorganisatie (bijvoorbeeld wel of geen hiërarchische structuur) De keuze voor een bepaalde evaluatiebenadering wordt uiteindelijk bepaald door opvattingen van docenten en bestuurder over de wijze waarop onderwijs functioneert.
Darling-Hammond, Wise en Pease (1983) noemen in navolging van Mitchell en Kerchner (1983), vier manieren waarop naar het onderwijsproces en met name naar de rol van de docent gekeken kan worden: "teaching work as labor, craft, profession or art".

"Teaching as labor".

Volgens deze visie is de docent verantwoordelijk voor het uitvoeren van een onderwijsprogramma dat door anderen ontworpen is. Hij/zij moet lesgeven op de in het programma voorgeschreven manier en daarbij rekening houden met de regels en voorschriften die binnen de school opgesteld zijn. De instelling verwacht niet van docenten dat zij op eigen initiatief wijzigingen in het programma aanbrengen. Docenten vormen als het ware een instrument voor de schoolleiding om bepaalde doelen op een zo efficiënt mogelijke wijze te bereiken. Men kan de docent binnen deze opvatting vergelijken met een werker in een productiebedrijf. De leiding formuleert de doelen, kiest een strategie, zorgt voor voldoende middelen en controleert of bepaalde nauw omschreven taken op correct wijze uitgevoerd worden.

"Teaching as craft".

In deze visie wordt de docent gezien als vakman: iemand die over de nodige specialistische kennis en technieken beschikt om zijn vak uit te oefenen en die zelfstandig in bepaalde situaties keuze's kan maken om de instelling voorgeschreven doelen te bereiken. De onderwijsorganisatie draagt er zorg voor dat docenten aangesteld worden die over voldoende algemeen- en vakdidactische kennis beschikken.
Onderwijs zoals dat vorm krijgt op middelbare scholen, ge-

tuigt in het algemeen van deze visie. De eindtermen van de opleiding zijn verwerkt in redelijk nauw omschreven eindexameneisen. Het is een taak van de docent dat leerlingen aan deze eisen kunnen voldoen, hetgeen impliceert dat de docent slechts binnen beperkte marges mogelijkheden heeft om leerstof te kiezen.

"Teaching as profession".

De docent wordt binnen deze visie gezien als een professional, die in principe het best geïnformeerd is over wat nodig is en hoe dat bereikt kan worden. Hij is nauw betrokken bij de opzet van het programma, of bepaalt zelf de inhoud van het programma. Hij bezit een aanzienlijke mate van zelfstandigheid. Van de docent wordt verwacht dat hij zowel over voldoende didactische kennis beschikt om een programma uit te voeren, als over voldoende inzicht om zelfstandig beslissingen te nemen over de inhoud en werkvormen binnen het programma. Het is de taak van de onderwijsorganisatie om didactisch en vakinhoudelijk bekwame, docenten aan te stellen en deze van voldoende middelen te voorzien om hun werk goed uit te voeren.

"Teaching as art".

In deze visie is de docent als creatief kunstenaar volledig verantwoordelijk voor de inhoud en uitvoering van zijn programma. Het geven van onderwijs bestaat niet uit het louter toepassen van beschikbare doceertechnieken, maar is mede afhankelijk van de inspiratie van de docent. Hoe het onderwijsproces verloopt, is in wezen onvoorspelbaar, zoals dat bij elk creatief proces in essentie het geval is. De kwaliteit van het onderwijs is in deze opvatting sterk afhankelijk van de persoon die doceert. De onderwijsorganisatie heeft slechts tot taak de voorwaarden te scheppen waaronder docenten hun werk in alle vrijheid verrichten.
Onderwijs zoals dat vorm krijgt op universiteiten, getuigt van deze opvatting. Docenten hebben een grote vrijheid wat betreft de keuze van leerstof, de manier waarop deze behandeld wordt, en de wijze waarop die getoetst wordt. Het ontbreken van eindtermen in de zin zoals die voor het basisonderwijs en middelbaar onderwijs vaak uitgeschreven zijn, leidt ertoe dat docenten in het universitaire onderwijs beduidend meer autonomie bezitten dan docenten in het basisonderwijs en middelbaar onderwijs.

De hierboven geschetste ideeën over onderwijs kunnen, net zoals bij de eerder genoemde evaluatiebenaderingen, onderverdeeld worden naar de mate waarin zij een rationalistische onderwijsaanpak veronderstellen. Volgens Darling-Hammond, Wise en Pease (1983) zijn deze ideeën zowel bepalend voor de manier waarop onderwijs aangepakt wordt als voor de manier waarop de onderwijsorganisatie gestructureerd wordt. Bij de

meer rationalistische ideeën over onderwijzen zal een groot gedeelte van de beslissingsbevoegdheid van docenten, ten aanzien van de keuze van doelstellingen, werkvormen, leerstofinhouden en toetsinstrumenten, aan hogere echelons in de onderwijsorganisatie (bijvoorbeeld faculteitsbestuur of onderwijsbeheerscommissie) worden voorbehouden. Genoemde auteurs maken onderscheid tussen twee typen schoolorganisaties: het rationalistische en het natuurlijke organisatiemodel. In tabel 1.2 zijn de kenmerken van deze organisatiestructuren weergegeven. In het eerste type is sprake van een hiërarchische opbouw, centrale besluitvorming, uitgewerkte regelgeving, en functionele scheiding in docentrollen (bijvoorbeeld de docent als toetsconstructeur, of de docent als onderwijsplanner). Bij een dergelijke organisatie past een evaluatie-benadering met een cursus- of docentoverstijgend karakter. Aangezien de belangrijkste beslissingen in bestuursorganen genomen worden, is informatie vereist die vooral geschikt is voor beleidsmatige kwesties. In het tweede type ligt de grootste verantwoordelijkheid voor het onderwijs bij de individuele docent. De docent bezit een grote autonomie en de organisatie creëert slechts een aantal randvoorwaarden om globale opleidingsdoelen te bereiken. In een dergelijke organisatie zou evaluatie vooral een bijdrage moeten leveren in de besluitvorming van individuele docenten. Hetgeen betekent dat de nadruk zou moeten liggen op beschrijving van het onderwijs in afzonderlijke cursussen.

Tabel 1.2 Kenmerken van twee typen schoolorganisaties, volgens Darling-Hammond, Wise en Pease (1983).

Rationalistische organisatiemodel
- hiërarchische opbouw van de organisatie - centrale besluitvorming - taakuitvoering a.d.h.v. voorgeschreven procedures om doelen te bereiken
Natuurlijke organisatiemodel
- afwezigheid van consensus over normen, waarden en onderwijsdoelen - organisatieopbouw rondom autonome groepen - decentrale besluitvorming - gebrekkige coördinatie met betrekking tot planning en besluitvorming.

Wat betekenen nu de hier geschetste opvattingen over het onderwijsproces en onderwijsorganisatie, voor de eerder gestelde vraag welke evaluatiebenadering het meest geschikt lijkt te zijn voor de medische faculteit? Zoals gezegd doen

Darling-Hammond et. al. (1983) een poging om een verband te leggen tussen concepties over het onderwijsproces enerzijds en evaluatiebenaderingen anderzijds. Doorslaggevend lijkt daarin te zijn de mate waarin docenten een opvatting over het onderwijsproces aanhangen die als rationalistisch aangemerkt zou kunnen worden. Dit heeft zowel consequenties voor de structuur van de onderwijsorganisatie als voor de betekenis die gehecht wordt aan een bepaalde evaluatiebenadering. In tabel 1.3 wordt aangegeven welke verbanden er volgens Darling-Hammond et. al. (1983) bestaan tussen de organisatievorm en de opvattingen over het onderwijsproces enerzijds en de na te streven evaluatiebenadering anderzijds. In onderwijsorganisaties die een meer rationalistisch karakter hebben en waarin docenten minder autonomie bezitten, zouden vooral evaluatiebenaderingen geschikt zijn die een meer modelmatig onderwijsproces veronderstellen. In onderwijsorganisaties waarin dit niet het geval is, lijkt vooral een aanpak als die van Parlett en Hamilton (1977) op zijn plaats.
Samenvattend kan gesteld worden dat een evaluator eerst moet beoordelen welke opvattingen betrokkenen hebben over onderwijs en hoe deze opvattingen uitgekristalliseerd zijn in de structuur van de onderwijsorganisatie, alvorens een evaluatiebenadering te kiezen.
In hoofdstuk 2 zal blijken dat binnen de medische faculteit gekozen is voor een evaluatiebenadering die uitgaat van de "models of teaching"-aanpak. Een benadering die, zoals zal blijken, aansluit bij het rationalistisch organisatiemodel van deze faculteit.

Tabel 1.3. Verbanden tussen organisatievormen, opvattingen over onderwijs en evaluatiebenaderingen.

Organisatievorm	Opvattingen over onderwijs	Evaluatiebenaderingen
Rationalistisch organisatie-model	Teaching as labor Teaching as craft	Doelstellingevaluatie Characteristics of teaching effectiveness Models of teaching
Natuurlijk organisatiemodel	Teaching as profession Teaching as art	Illuminatieve evaluatie

1.3 Methoden van dataverzameling in programma-evaluatie.

Informatie over de kwaliteit van het onderwijs kan op verschillende manieren en bij verschillende bronnen verzameld worden. Alom bekende onderzoeksmethoden uit de sociale wetenschappen worden ook in evaluatie-onderzoek gebruikt: observaties, interviews, en vragenlijsten. In deze paragraaf wordt achtereenvolgens aandacht besteed aan de volgende informatiebronnen: de student, de collega-docenten, oud-studenten, de docent zelf en externe observatoren. We zullen vooral aandacht besteden aan studentoordelen als indicator van de onderwijskwaliteit, omdat het in dit proefschrift beschreven evaluatiesysteem voornamelijk gebruik maakt van deze methode. Van de andere methoden wordt slechts in het kort aangeven wat de voor- en nadelen zijn.

1.3.1 Studenten

Studenten vormen de meest gebruikte en tevens de belangrijkste bron van informatie t.b.v. evaluatie in het hoger onderwijs, omdat zij, als ontvangers van onderwijs, in een bijzonder goede positie zijn om een oordeel te geven over het functioneren van een cursus. Bovendien kan men effecten van onderwijsprogramma's meten door studenten toetsen te laten maken die de leerstof in een cursus weerspiegelen.

Studentoordelen.

Studentoordelen worden het meeste gebruikt om het onderwijs te evalueren. Dat is begrijpelijk want ze kunnen gezien worden als een vorm van observaties, door studenten verricht tijdens een cursus. Die oordelen worden verzameld met behulp van beoordelingsvragenlijsten.
Deze lijsten worden, meestal aan het einde van een cursus, aan studenten voorgelegd en bevatten in de meeste gevallen twintig tot zestig items over het functioneren van de docent, de organisatie van de cursus, de leerstofinhoud, etc..
Studentoordelen hebben een aantal voordelen die door Costin, Greenough en Menges (1971) als volgt omschreven worden:

1. "Such ratings could provide feedback which the instructor might not be able to elicit from students on a face-to-face basis.
   (This information alone, with no sanctions contingent, could improve teaching).
2. They could provide departmental and college-wide norms against which individual faculty ratings could be judged.
3. They could provide a way in which a faculty member could, if he desired, demonstrate his undergraduate teaching effectiveness to those who have expressed an interest in evaluating this parameter for salary increase.

4. They could provide information to the department and college on areas of relative strenght or weakness in undergraduate teaching, suggest directions for the development of new courses or programs, and provide evaluative information and norms on the various new programs which are implemented.
5. They could provide the student with a source of information to aid him in the selection of courses."
(Dit laatste is vooral in de Verenigde Staten van belang en kan gezien het hoger onderwijsbeleid van de overheid ook hier ten lande in belang toenemen).

Desondanks zijn aan studentoordelen ook een aantal problemen verbonden. Docenten en andere belanghebbenden koesteren soms ernstige reserves ten aanzien van de betrouwbaarheid en validiteit van studentoordelen. Men betwijfelt dan met name of studenten wel over voldoende expertise en ervaring beschikken om het onderwijs op waarde te schatten, en in het bijzonder gelooft men dat studenten in hun oordeel beïnvloed worden door factoren die niet gerelateerd zijn aan de kwaliteit van het onderwijs (Page, 1974; Dowell & Neal, 1982).
Er is zeer veel onderzoek verricht naar dit laatste punt. Vanuit de veronderstelling dat wie een hoger cijfer behaalde op een tentamen een cursus wel beter zou beoordelen, is er bijvoorbeeld onderzoek verricht naar de samenhang tussen tentamenresultaten en de hoogte van de beoordeling (Feldman, 1976; Howard & Maxwell, 1980; Vollmer, 1986).
Andere onderzoeken richten zich op de invloed van de aantrekkelijkheid van het vak, of het onderhoudende, aardige optredende docent op de beoordeling van de cursus. Dit vanuit de veronderstelling dat als studenten iets leuk of interessant vinden, ze eerder geneigd zijn om positievere oordelen te geven (zie bijvoorbeeld Marsh, 1984). Alles bij elkaar genomen kan gesteld worden dat de meeste onderzoeken tot de conclusie leiden dat studentoordelen voldoende betrouwbaar en valide zijn om voor beoordelingsdoeleinden gebruikt te worden (Cohen, 1980; Howard & Conway, 1985). In de hoofdstukken 3 en 4 zullen we nader op deze problematiek ingaan.
Een ander probleem betreft het soort informatie dat men verkrijgt met behulp van beoordelingsvragenlijsten. In het hoger onderwijs in de Verenigde Staten vervullen studentoordelen binnen faculteiten vaak twee functies, namelijk: het geven van feedback aan docenten gericht op het verbeteren van cursussen, en het geven van informatie aan bestuurders over het functioneren van docenten om beslissingen t.o.v. promotie e.d. te ondersteunen met feitelijke gegevens. Om beide functies te vervullen zouden vragenlijsten items moeten bevatten die zowel relevante informatie geven aan docenten als aan degenen die beslissingen nemen over docenten. In de praktijk blijkt dit echter op problemen te stuiten. Rotem en Glasman (1979) verwoordden deze problemen als volgt: 'A potential drawback of students' ratings is that the feedback they produce may be too general to provide guidance to teachers.

Students' ratings seem to be attractive because of their numerical properties which make them efficient and easy to score. But this attribute may also be the source of their limitation in providing diagnostic feedback to teachers (not unlike the limitations of grades given to students, which tell only whether one has done well or not). Some students' ratings may not be sufficient to reveal what particular aspects of teaching should be modified and by what means.'

Uit onderzoek blijkt dat docenten vooral behoefte hebben aan cursusspecifieke evaluatie en niet aan standaardevaluaties die in alle cursussen gebruikt kunnen worden (Brandenburg, Nassauer & Buckmaster, 1981). Facultaire bestuurders daarentegen, willen het tegenovergestelde: informatie die het hun mogelijk maakt vergelijkingen tussen cursussen te maken. In de praktijk ziet men dan ook dat, afhankelijk van degene die de evaluatievraag stelt, het evaluatiesysteem een specifieke vorm heeft. Men heeft evaluatiesystemen ontwikkeld waarin een voor iedere cursus specifieke, vragenlijst gemaakt wordt (het zogenaamde cafetaria- of zelfbedieningssysteem), systemen die uitsluitend standaardvragenlijsten gebruiken en systemen die een aantal standaarditems combineren met cursusspecifieke items.
Een voorbeeld van een gecombineerd systeem is Students' Evaluations of Educational Quality. Dit zogenaamde SEEQ-systeem is ontwikkeld door een groep onderwijskundige onderzoekers aan de Universiteit van California te Los Angeles (Marsh, 1982a), vormt een compromis tussen een systeem met uniforme vragenlijsten en een aanpak berustend op flexibele, op de afzonderlijke cursussen toegesneden, vragenlijsten. SEEQ bestaat uit een vragenlijst waarin dertig standaardvragen zijn opgenomen over het functioneren van de docent, de opzet van de cursus, de leerstof en het tentamen. Docenten kunnen zelf naar wens een aantal vragen toevoegen die specifiek zijn voor hun cursus. De evaluatieresultaten worden toegezonden naar de betrokken docent en naar de dekaan van de faculteit. De resultaten worden dus gebruikt zowel voor het verbeteren van de cursussen als voor het nemen van personele beslissingen over het docentencorps. Jaarlijks wordt voor de studenten een overzicht gemaakt van de evaluatieverslagen. Studenten kunnen dit overzicht gebruiken voor het selecteren van cursussen. Op wens van de faculteit kunnen aparte analyses verricht worden om bijvoorbeeld het effect van allerlei innovaties in het onderwijsprogramma te onderzoeken. Doordat de vragenlijst een aantal standaardvragen bevat, kan het faculteitsbestuur cursussen op een aantal punten met elkaar vergelijken. In onderstaande tabel worden bij wijze van voorbeeld enkele vragen gegeven uit de SEEQ.

Tabel 1.4 Enkele items uit de SEEQ (Marsh, 1982a)

		Very Poor	Poor	Moder-ate	Good	Very Good
1 Learning:	You found the course intellectually challenging and stimulating.	1	2	3	4	5
2	You have learned something which you consider valuable.	1	2	3	4	5
3 Enthusiasm:	Instructor was enthusiastic about teaching in the course.	1	2	3	4	5
4	Instructor enhanced presentations with the use of humour.	1	2	3	4	5
5 Organisation:	Instructor's explanations were clear.	1	2	3	4	5

Een voorbeeld van een cursusspecifiek evaluatiesysteem is het Nederlandse ISEK-systeem (Instrumentarium voor Systematische Evaluatie van Cursussen). Dit systeem is ontwikkeld door het Centrum voor Onderzoek van het Wetenschappelijk Onderwijs van de Rijksuniversiteit Groningen (Bartelds, Joostens & Kluiter, 1983). Voor elke te onderzoeken cursus wordt een nieuw evaluatie-instrument ontwikkeld. Docenten kunnen uit een aantal instrumenten kiezen om hun cursus te evalueren: een vragenlijst, interviews met studenten of panelbijeenkomsten met docenten en studenten. Als docenten een keuze maken voor de vragenlijst als methode kunnen zij uit een catalogus items selecteren. De vragenlijst bevat geen items die standaard opgenomen moeten worden voor iedere cursus. In tegenstelling tot de SEEQ worden bij het ISEK evaluatieresultaten alleen gebruikt voor het verbeteren van cursussen.

Interviews.

Interviews met studenten zijn met name geschikt als onderzoekers aanvullende informatie willen over een cursus. Naar aanleiding van de uitkomsten verkregen met vragenlijsten, kan men besluiten om groepen studenten te vragen gedetailleerde beschrijvingen te geven van hun ervaringen met een cursus. Een nadeel van deze methode is dat het een groot beslag legt op de tijd van studenten en docenten.
Deze methode wordt dan ook niet standaard gebruikt maar slechts als er sprake is van problemen met een cursus, zonder dat de precieze aard van die problemen duidelijk is.

Toetsresultaten.

Leerprestaties kunnen, zoals we in de vorige paragraaf gezien hebben, een belangrijk criterium vormen om de kwaliteit van het onderwijs te bepalen. Een probleem is echter dat de uitslag van een toets geen maat is voor de aktiviteiten, inzet en kwaliteit van de docent, maar in principe voor de aktiviteiten van de student. Uit onderzoek is gebleken dat toetsresultaten slechts een gedeeltelijke weerspiegeling vormen van de moeite die de docent zich in z'n onderwijs getroost heeft (Cohen, 1981; Marsh, 1984). Een ander probleem is dat deze indicator nauwelijks informatie geeft over noodzakelijke wijzigingen in een programma. Een toetsuitslag als zodanig geeft immers geen indicatie over de eventuele knelpunten die zich hebben voorgedaan tijdens de cursus. In combinatie met andere indicatoren (interviews met studenten, studentoordelen) kunnen toetsresultaten echter soms wel nuttige informatie geven, met name bij de beantwoording van de vraag in hoeverre bepaalde doelstellingen in het onderwijs ook werkelijk gerealiseerd zijn.

1.3.2 Collega-docenten (peers)

Collega-docenten kan gevraagd worden om een docent te beoordelen, de zogenaamde "peer-review". Braskamp, Brandenburg en Ory (1984) noemen de volgende voordelen van deze methode:
- collega-docenten zijn didactische experts op hetzelfde vakgebied,
- collega-docenten zijn in staat om de doelstellingen van een cursus op hun adequaatheid te beoordelen,
- collega-docenten kunnen beoordelen of alternatieve werkvormen meer geschikt zouden zijn.

Ondanks deze op het oog aantrekkelijke voordelen, wordt deze methode weinig toegepast. Marsh (1984) merkt op dat aan deze methode een aantal ernstige nadelen kleven: 1) oordelen van collega-docenten blijken minder betrouwbaar en valide te zijn dan andere vormen van beoordeling, 2) ze zijn meer bedreigend voor docenten, 3) ze worden meer beïnvloed door factoren die weinig met onderwijs te maken hebben, zoals de onderzoeksproductiviteit van de beoordeelde docent.

1.3.2 Docenten.

Zelf-beoordelingen van docenten.

Deze methode wordt voornamelijk gebruikt om studentoordelen te valideren. Uit onderzoek blijkt dat zelf-oordelen van docenten zwak tot matig gecorreleerd zijn met studentoordelen. De gevonden waarden variëren tussen .20 en .50. Om zelfbeoordelingen van docenten te verzamelen maakt men meestal gebruik van dezelfde vragenlijsten die ook voor studenten gebruikt worden, zij het in een aangepaste vorm. Een probleem met zelfbeoordelingen wordt gevormd door de vraag welke in-

formatie zij bieden. Met andere woorden: hoe valide zijn die oordelen? De omstandigheid dat bepaalde antwoorden op vragen sociaal gewenst zijn, vormt meestal een bedreiging voor de validiteit. Immers, welke docenten zijn genegen om zichzelf een slechte beoordeling te geven als deze informatie verzameld wordt door derden?

Onderzoeksproductiviteit van docenten.

Deze indicator wordt als zodanig nauwelijks gebruikt om de kwaliteit van het onderwijs te meten. Desondanks wordt hij wel eens betrokken in discussies over de relatie tussen onderzoek en onderwijs. Onderwijs geven en onderzoek verrichten vormen de twee belangrijkste taken voor docenten in het universitaire onderwijs. Het verrichten van onderzoek helpt docenten om op de hoogte te blijven van de nieuwste ontwikkelingen op hun vakgebied, hetgeen de kwaliteit van het onderwijs zou bevorderden. Deze gedachtengang wordt door sommigen vaak aangegrepen om te voorkomen dat er systematische evaluatie van het onderwijs plaatsvindt. Uit onderzoek is echter gebleken dat er geen correlatie bestaat tussen de onderzoeksproductiviteit (artikelen in wetenschappelijke tijdschriften, interne en externe rapporten, etc.) van een docent en de beoordeling die hij krijgt van studenten (Marsh, 1984).

1.3.4 Oud-studenten.

Oud-studenten kunnen bruikbare informatie leveren over de kwaliteit van het onderwijs als het gaat om de vraag in hoeverre de opleiding voldoende aansloot op de behoeften van de arbeidsmarkt. Gezien het grote tijdsinterval dat vaak ontstaat tussen het moment waarop informatie verkregen wordt en het moment waarop het betreffende onderwijs beëindigd werd, is deze informatie weinig bruikbaar voor de verbetering van lopende cursussen.
Deze methode wordt nauwelijks toegepast als de evaluatie gericht is op het verbeteren van het onderwijsproces. Ze wordt wel gebruikt als de evaluatie gericht is op het bepalen van de adequaatheid van het product van de opleiding en van de opleidingsdoelen.

1.3.5 Externe observatoren.

Externe observatoren zijn personen die speciaal getraind worden om docentengedrag te observeren. Deze personen zijn niet persoonlijk betrokken bij de cursus. De observaties zijn gericht op het registreren van zo concreet mogelijke docentgedrag (bijvoorbeeld, het aantal vragen dat de docent tijdens een lesuur aan studenten stelt).
Deze methode wordt voornamelijk in het lager onderwijs toegepast. Uit onderzoek in het hoger onderwijs blijkt dat de betrouwbaarheid en validiteit van dit soort observaties redelijk hoog is. Marsh (1984) noemt een aantal studies waarin

substantiële correlaties gevonden werden tussen observaties van docentgedrag en de tentamenprestaties van studenten.

1.4 Het gebruik van gegevens uit programma-evaluaties.

Evaluatie van onderwijsprogramma's kan gericht zijn op het verschaffen van informatie ten behoeve van verbetering (de zogenaamde formatieve functie), of op het verschaffen van informatie aan derden (bijvoorbeeld faculteitsbestuur of overheid) teneinde verantwoording aan deze af te leggen (de zogenaamde summatieve functie). Dit onderscheid vertoont sterke overeenkomsten met het in de eerste paragraaf genoemde onderscheid tussen interne en externe kwaliteitsbewaking. Als evaluatie een formatieve functie vervult, moet de informatie bruikbaar zijn om een bijdrage te leveren aan de verandering of verbetering van het onderwijs in een cursus of programma. Het kan dan bijvoorbeeld informatie betreffen over de aansluiting van het onderwijs op de voorkennis van studenten, informatie over het gedrag van docenten, informatie over het gebruik van leermiddelen, etc.. Als evaluatie een summatieve functie heeft, dient informatie verschaft te worden waardoor bijvoorbeeld beslissingen genomen kunnen worden over het al dan niet voortzetten van een programma, de aanstelling van docenten, etc.. Een voor de hand liggende vraag is in hoeverre evaluatieresultaten werkelijk een rol spelen bij het nemen van beslissingen die verbeteringen beogen, of bij het nemen van beslissingen over het voortzetten van opleidingen.
In paragraaf 1.2 is gebleken dat de meest gangbare evaluatiebenaderingen, namelijk de "characteristics of teaching"-benadering en de "models of teaching"-benadering, gericht zijn op de identificatie van proceskenmerken die het onderwijsproduct in gunstige zin beïnvloeden. De vraag hoe de verkregen informatie gebruikt moet worden, wordt door vertegenwoordigers van deze stromingen zelden gesteld. Men veronderstelt, min of meer stilzwijgend, dat docenten of organisaties, de verstrekte informatie vanzelf zullen gebruiken om veranderingen aan te brengen in het onderwijs. Met andere woorden, de presentatie van evaluatiegegevens zou voldoende garantie bieden dat daadwerkelijk veranderingen in het onderwijs plaatsvinden. In de praktijk blijkt echter dat onderwijsevaluatie niet automatisch leidt tot kwaliteitsverbetering (Cooley & Bickel, 1986). Integendeel, het nemen van beslissingen over veranderingen in het onderwijs wordt door meer factoren beïnvloed dan alleen de uitkomsten van evaluatie. Uit onderzoek is bijvoorbeeld gebleken dat beslissingen over veranderingen in een programma eerder tot stand komen als de evaluatie-uitkomsten stroken met de verwachtingen of percepties van bestuurders. Als de uitkomsten daarmee strijdig zijn, neemt de kans af dat het programma veranderd wordt (Cousins & Leithwood, 1986). Slagen of falen van evaluatie is met andere woorden helaas niet alleen afhankelijk van de kwaliteit van de geboden informatie zelf, maar ook van de mate waarin potentiële gebruikers ook werkelijk door de

informatie bereikt worden. In deze paragraaf zullen we daarom aandacht besteden aan de problematiek rond het gebruik van evaluatie-uitkomsten. In hoofdstuk 5 van dit proefschrift zal onderzoek gepresenteerd worden over de wijze waarop informatie, verkregen uit evaluaties, in de medische faculteit van de Rijksuniversiteit Limburg gebruikt wordt.

Onderzoek naar het gebruik van evaluatiegegevens kan op twee verschillende vragen betrekking hebben. In de eerste plaats kan het onderzoek gericht zijn op de vraag of de gebruiker zijn beslissingen al dan niet baseert op de aangereikte informatie. De vraag is dan in hoeverre de gebruiker bij het nemen van beslissingen feitelijk beïnvloed wordt door evaluatieresultaten. In de tweede plaats kan het gericht zijn op de vraag wat de effecten zijn van beslissingen die gebaseerd zijn op evaluatieresultaten. De vraag is in dit geval of evaluatie daadwerkelijk tot verbeteringen of veranderingen in het onderwijs leidt. In het onderstaande gedeelte worden achtereenvolgens beide vragen behandeld.

1.4.1 Instrumenteel versus conceptueel gebruik van programma-evaluatiegegevens.

Cooley en Bickel (1986) maken een onderscheid tussen instrumenteel en conceptueel gebruik van evaluatieresultaten. Instrumenteel gebruik van de resultaten van een evaluatie verwijst naar situaties waarin die resultaten expliciet betrokken worden in de besluitvorming over een bepaald programma. Met conceptueel gebruik wordt bedoeld dat programma-evaluatie van invloed kan zijn op de gedachtengang van personen die geacht worden beslissingen te nemen zonder dat een beslissing het directe resultaat hoeft te zijn van de gepresenteerde evaluatieve informatie. Onderzoek naar het feitelijk gebruik van evaluatieresultaten heeft zich met name gericht op de vraag welke factoren dat gebruik in positieve of negatieve zin beïnvloeden (Levinton & Hughes, 1981; Cousins & Leitwood, 1986; Cooley & Bickel, 1986). Levinton en Hughes (1981) noemen vijf factoren die het gebruik van evaluatieresultaten beïnvloeden: de relevantie van de informatie, de wijze waarop de communicatie tussen evaluator en gebruiker verloopt, de manier waarop de informatie gepresenteerd wordt, de geloofwaardigheid van de informatie, en de betrokkenheid van de gebruiker bij de evaluatie. In tabel 1.5 zijn die factoren met steekwoorden aangeduid. We zullen de betreffende factoren allen kort bespreken.

Tabel 1.5: Factoren die het gebruik van informatie beïnvloeden.

---

1. RELEVANTIE.
   - aansluiting van de evaluatie op de behoeften van de gebruiker,
   - tijdsplanning van het evaluatie-onderzoek: tijdstip van presentatie.

2. COMMUNICATIE.
   - directheid van de communicatie tussen gebruiker en evaluator,
   - verspreiding van de informatie.

3. PRESENTATIE.
   - presentatie van informatie,
   - mate van begrip van de gepresenteerde informatie bij de gebruiker.

4. GELOOFWAARDIGHEID.
   - overeenstemming met andere informatiebronnen,
   - vooropgezette meningen van de gebruiker ten aanzien van de bruikbaarheid van het onderzoek in het algemeen,
   - geloofwaardigheid van de evaluator,
   - kwaliteit van het onderzoek.

5. BETROKKENHEID VAN DE GEBRUIKER.
   - persoonlijke betrokkenheid met het evaluatie-onderzoek.

De eerste factor, relevantie, heeft betrekking op de informatiebehoeften van de gebruiker en op het tijdstip van presentatie van informatie. Naarmate informatie verschaft wordt die inhoudelijk aansluit op de informatiebehoeften van de gebruiker (zie par. 1.2) en naarmate deze informatie tijdig wordt gepresenteerd, zal deze eerder gebruikt worden. De tweede factor, communicatie, heeft betrekking op de contacten tussen de evaluator en de gebruiker van de informatie. Directe communicatie, dat wil zeggen zonder tussenkomst van derden, is positief gecorreleerd met het gebruik van evaluatieresultaten. Volgens Levinton en Hughes (1981) is daarbij de aard van de communicatie (schriftelijk of mondeling) niet van belang, maar wel de frequentie en de directheid.
De derde factor, presentatie, heeft betrekking op de helderheid van de presentatie (grafieken, commentaren, al dan niet gebruik van jargon). Deze factor blijkt een positieve invloed te hebben op het gebruik van informatie.
De vierde factor, geloofwaardigheid van het evaluatie-onderzoek zoals gepercipieerd door gebruikers van informatie, slaat op de vooropgezette meningen van gebruikers ten aanzien van de kwaliteit van het onderzoek, de geloofwaardigheid van

de onderzoeker(s), de kwaliteit van de gevonden resultaten, en de relatie van deze resultaten met andere reeds bij de gebruiker aanwezige informatie.
Volgens Levinton en Hughes (1981) is gebleken dat het gebruik van evaluatieresultaten positief beïnvloed wordt als deze overeenkomen met de verwachtingen en percepties van de gebruiker, als de gebruiker meer vertrouwen in het onderzoek heeft dan in zijn eigen intuïtieve oordelen, en als de toegepaste onderzoeksmethodologie niet makkelijk aanvechtbaar is. Gebruik wordt negatief beïnvloed als de resultaten strijdig zijn met de verwachtingen van de gebruiker, en als de gebruiker de kwaliteit van het onderzoek laag inschat.
De vijfde factor betreft de betrokkenheid van de gebruiker bij het evaluatie-onderzoek. Naarmate de gebruiker vanaf het begin meer betrokken is geweest bij de opzet en uitvoering van het onderzoek, neemt ook het gebruik van informatie toe.

1.4.2 Bruikbaarheid en effectiviteit van studentoordelen.

Tot nog toe is uitsluitend gesproken over de vraag welke factoren een rol spelen bij het gebruik van evaluatieresulaten. De vraag resteert in hoeverre studentoordelen door docenten gebruikt worden bij het nemen van beslissingen over het onderwijs en welke effecten deze beslissingen hebben op de kwaliteit van het onderwijs.
Met "Bruikbaarheid van studentoordelen" kunnen twee zaken bedoeld worden: de subjectieve en de objectieve bruikbaarheid van gegevens (Rotem & Glasman, 1979; Wilson, 1986). Voor de subjectieve bruikbaarheid geldt de gebruiker als criterium. Is hij van mening dat de gegevens nuttig zijn en als basis voor verandering kunnen dienen? In het geval van objectieve bruikbaarheid wordt bruikbaarheid opgevat als "resulterend in feitelijk positieve effecten op het onderwijs". Om verwarring te voorkomen, wordt het volgende onderscheid gemaakt bij de invulling van het begrip bruikbaarheid. Als gesproken wordt over de <u>bruikbaarheid</u> van informatie of studentoordelen, dan wordt naar de eerste betekenis verwezen. Als gesproken wordt over het <u>effect</u> van informatie of studentoordelen, dan wordt de tweede betekenis bedoeld.
De bruikbaarheid van studentoordelen is voornamelijk onderzocht door er gebruikers naar te vragen (Brandenburg, Nassauer & Buckmaster, 1981; Ory & Braskamp, 1981; Wilson, 1986). Brandenburg et. al. (1981) melden dat docenten zich vooral bekommeren over de specificiteit van de items in beoordelingsvragenlijsten. Men heeft vooral behoefte aan cursusspecifieke informatie; in standaardinformatie die in iedere cursus op dezelfde wijze verzameld is, is men minder geïnteresseerd. Het doel waarvoor studentoordelen gebruikt worden, vormt een probleem voor docenten. Men is van mening dat studentoordelen redelijk bruikbaar zijn als het erom gaat aspecten van het gegeven onderwijs te verbeteren, maar niet als het erom gaat informatie aan te dragen ten behoeve van het nemen van personele beslissingen, bijvoorbeeld het omzetten

van een tijdelijk dienstverband in een vaste aanstelling, of het bevorderen naar een hogere salarisschaal.
Ory en Braskamp (1981) onderzochten aan welke vorm van feedback docenten het meeste waarde hechten. In hun onderzoek werden aan docenten drie verschillende, op fictieve gegevens gebaseerde, evaluatierapporten voorgelegd. In deze rapporten werden gegevens gepresenteerd die verkregen waren met verschillende dataverzamelingsmethoden: interviews, beoordelingsvragenlijsten met uitsluitend gesloten vragen en beoordelingsvragenlijsten met uitsluitend open vragen. Aan docenten werd gevraagd de rapporten op de volgende punten te beoordelen: accuraatheid, geloofwaardigheid, bruikbaarheid, interpreteerbaarheid, uitgebreidheid, betrouwbaarheid en waardevolheid. Uit dit onderzoek blijkt dat de informatiewaarde van een evaluatierapport in het algemeen als groter beoordeeld wordt, naarmate het zich op specifieke onderdelen van het beoordeelde onderwijs richt en aanwijzingen voor kwaliteitsverbetering bevat. Docenten hebben de meeste waardering voor verslagen van groepsinterviews, vanwege de uitgebreidheid van de informatie. Docenten vinden antwoorden op open vragen minder betrouwbaar dan op gesloten vragen en interviews.
Onderzoek naar de effecten van studentoordelen op de kwaliteit van het onderwijs heeft nogal tegenstrijdige resultaten opgeleverd. Volgens sommige onderzoekers heeft programma-evaluatie nauwelijks een effect op de kwaliteit van het onderwijs (Abrami, Leventhal & Perry, 1979). Andere onderzoekers zijn minder pessimistisch. Cohen (1980) verrichtte een meta-analyse op 22 studies waarin de effecten van studentoordelen onderzocht werden. De belangrijkste conclusie was dat evaluatieresultaten voornamelijk effect sorteren als schriftelijke rapportage gepaard gaat met persoonlijke contacten tussen evaluator en docent. Centra (1973) vond vergelijkbare resultaten. Bij docenten die uitsluitend schriftelijke informatie kregen over hun cursus, bleken nauwelijks verbeteringen in het onderwijs op te treden.

1.5 Normering.

In de voorgaande paragrafen werd achtereenvolgens aandacht besteed aan de problematiek van het definiëren van "kwaliteit van het onderwijs", aan enkele problemen rond de meetbaarheid van onderwijskwaliteit, en aan vragen naar het gebruik en de effecten van evaluatieresultaten. De vraag in welke gevallen, namelijk bij welke met het meetinstrument verkregen waarden, sprake is van voldoende of onvoldoende kwaliteit, werd tot nog toe buiten beschouwing gelaten. Dit normeringsvraagstuk komt in deze paragraaf kort aan bod.
In de literatuur worden drie oplossingen beschreven voor het normeringsprobleem in evaluatie-onderzoek: 1) absolute normering, 2) relatieve normering, en 3) zelfnormering (Bloom, 1970; Marsh, 1982; Bartelds, Joostens & Kluijter, 1983). Absolute normering wordt vooral gepropageerd door vertegen-

woordigers van de doelstellinggeoriënteerde evaluatiebenadering (Tyler, 1950; Bloom, 1970). De mate waarin de doelstellingen van een onderwijsprogramma bereikt zijn, bepalen volgens deze auteurs het oordeel over de kwaliteit van het onderwijsprogramma. Als bijvoorbeeld de doelstellingen voor 80% of meer bereikt zijn, spreekt men van een goed onderwijsprogramma. De normen waaraan een programma moet voldoen, worden vooraf gesteld.
Relatieve normering wordt bijvoorbeeld toegepast binnen het SEEQ-systeem. Om te bepalen of cursussen goed of slecht gefunctioneerd hebben, worden vergelijkingen tussen cursussen gemaakt. Daartoe worden op een aantal factoren die aan de vragenlijst ten grondslag liggen, scores berekend die als uitgangspunt dienen voor een onderlinge rangordening van de geëvalueerde cursussen. Binnen dit systeem is het gebruikelijk dat ook docenten ten opzichte van elkaar gerangschikt worden (Marsh, 1982). Na afloop van een cursus ontvangen docenten een rapport waarin per item de percentuele rangscore wordt weergegeven.
Het ISEK-systeem gaat uit van een andere benadering (Joostens, Bartelds & Kluijter, 1983). Aangezien dit systeem cursusspecifiek is en men niet altijd gebruik maakt van een vragenlijst, kunnen er geen systematische vergelijkingen tussen alle cursussen van het onderwijsprogramma plaatsvinden. Als docenten kiezen voor de vragenlijst als evaluatie-instrument, worden zij verzocht om per item een zogenaamd grenspercentage aan te geven. Onder dit percentage verstaat men het percentage antwoorden dat voor de docent de grens markeert tussen problematische en niet-problematische evaluatie-uitkomsten. De docenten bepalen dus binnen dit systeem, in tegenstelling tot het SEEQ, zelf de normen waaraan de kwaliteit van het onderwijs zal worden afgemeten.
Het normeringsvraagstuk zal in hoofdstuk 5 verder ter sprake komen.

1.6 Het thema van dit proefschrift.

In dit proefschrift staat de betrouwbaarheid, validiteit en bruikbaarheid van studentoordelen centraal. Studentoordelen vormen een essentieel onderdeel van het systeem voor interne kwaliteitsbewaking van de medische faculteit van de Rijksuniversiteit Limburg. Deze methode van dataverzameling wordt in deze faculteit gebruikt om de kwaliteit van het onderwijsprogramma te meten en te verbeteren.
Centraal thema in dit proefschrift is dan ook de vraag in hoeverre studentoordelen deze rol, binnen het kader van interne kwaliteitsbewaking, kunnen vervullen. Daarbij gaat het enerzijds om de vraag of studentoordelen voldoende psychometrische kwaliteiten (betrouwbaarheid en validiteit) bezitten om kwaliteit van het onderwijs te meten, en anderzijds om de vraag of studentoordelen gebruikt kunnen worden om de kwaliteit van het onderwijs te verbeteren. Alvorens een overzicht van deze hoofdstukken te geven, wordt nu samenvattend het

globaal overzicht afgerond van problemen die zich kunnen voordoen bij de evaluatie van universitair onderwijs.
Deze problemen hebben achtereenvolgens betrekking op: 1) de formulering van beoordelingscriteria, 2) de keuze van meetinstrumenten, 3) het feitelijk gebruik van evaluatieresultaten, en 4) het vraagstuk van de normering.
In het begin van dit hoofdstuk is uitvoerig stilgestaan bij het eerste probleem. Aan de hand van een bespreking van vier gangbare evaluatiebenaderingen werden verschillende strategiën gepresenteerd voor de formulering van criteria. Uit deze bespreking kwam naar voren dat er geen eenduidige beoordelingscriteria bestaan. Met andere woorden, de kwaliteit van het onderwijs bestaat niet. Vandaar dat tussen evaluatiebenaderingen aanzienlijke verschillen kunnen bestaan wat betreft de elementen van het onderwijs die onderzocht worden.
Daarna werden enkele methoden beschreven voor de verzameling van data die een valide en betrouwbare operationalisatie moeten vormen van de gekozen beoordelingscriteria. Uit paragraaf 1.3 bleek dat het verzamelen van student- oordelen met behulp van beoordelingsvragenlijsten, de meest gebruikte methode is. Aan deze methode zijn een aantal voor de hand liggende voordelen verbonden: ze is relatief goedkoop, weinig arbeidsintensief, en ze lijkt betrouwbaar en valide te zijn.
In paragraaf 1.4 werd aandacht besteed aan het feitelijk gebruik van evaluatieresultaten en aan de mogelijke effecten van evaluatieresultaten op veranderingen in de kwaliteit van het geëvalueerde onderwijs. In deze paragraaf werd geconcludeerd dat docenten van mening zijn dat studentoordelen voldoende informatie geven over het functioneren van het onderwijs. Bovendien bleek dat evaluatieresultaten voornamelijk effectief lijken te zijn als schriftelijke rapportage gepaard gaat met persoonlijke contacten tussen evaluator en gebruiker. Tenslotte werden in paragraaf 1.5 enkele normeringsmethoden besproken.
Hoofdstuk 2 behandelt de constructie van een meetinstrument dat binnen de medische faculteit ontwikkeld is voor de evaluatie van het onderwijsprogramma. Daarbij wordt achtereenvolgens aandacht besteed aan constructiemethoden voor beoordelingsvragenlijsten, aan de onderwijskundige uitgangspunten die aan het onderwijsprogramma van de medische faculteit ten grondslag liggen en aan enkele onderwijsleermodellen die aanknopingspunten bieden voor een theoretische onderbouwing van het instrument.
Hoofdstuk 3 geeft een overzicht van de literatuur die betrekking heeft op de psychometrische kwaliteit van studentoordelen. Daarin wordt een conceptueel model geschetst dat als basis kan dienen voor onderzoek naar de betrouwbaarheid en validiteit van studentoordelen. Tevens wordt een overzicht gegeven van onderzoeksbevindingen uit de literatuur met betrekking tot de betrouwbaarheid en validiteit van studentoordelen.
Hoofdstuk 4 beschrijft empirisch onderzoek naar de betrouwbaarheid en validiteit van studentoordelen die binnen de

medische faculteit verzameld worden met behulp van het in hoofdstuk 2 beschreven instrument.
Hoofdstuk 5 tenslotte, bevat vier empirische studies naar het gebruik van studentoordelen. In deze studies wordt getoond hoe studentoordelen gebruikt kunnen worden om de kwaliteit van het onderwijs te verbeteren.

HOOFDSTUK 2. BESCHRIJVING VAN HET EVALUATIESYSTEEM VOOR HET CURRICULUM VAN DE MEDISCHE FACULTEIT.

2.1 Inleiding.

In dit hoofdstuk volgt een beschrijving van de evaluatiebenadering die aan de medische faculteit van de Rijksuniversiteit Limburg gebruikt wordt. Deze benadering werd in 1981, binnen de vakgroep Onderwijsontwikkeling en Onderwijsresearch, ontwikkeld om bij het onderwijs betrokkenen van informatie te voorzien teneinde de kwaliteit van het onderwijs te verbeteren. Kenmerkend voor deze benadering is dat het een compromis vormt tussen zogenaamde cursusspecifieke en standaard evaluatiesystemen. Qua opzet lijkt deze benadering sterk op het in hoofdstuk 1 beschreven SEEQ-systeem. Als voornaamste dataverzamelingsmethode wordt, evenals bij het SEEQ-systeem, gebruik gemaakt van beoordelingsvragenlijsten. Voordat we aandacht besteden aan de vragenlijst zelf, beschrijven we eerst in paragraaf 2.2 de uitgangspunten van het evaluatiesysteem van de medische faculteit.
Vervolgens worden in 2.3 in het kader van instrumentontwikkeling een aantal methoden beschreven om beoordelingsvragenlijsten te construeren. Daarna volgt de beschrijving achtereenvolgens de historie van probleemgestuurd medisch onderwijs (paragraaf 2.4), het onderwijsprogramma van de medische faculteit (paragraaf 2.5) en enkele belangrijke theorieën over onderwijs (paragraaf 2.6), op grond waarvan vragen ten behoeve van de vragenlijsten geconstrueerd zijn. Tevens wordt relevant onderzoek naar kenmerken van probleemgestuurd medisch onderwijs besproken. Uit deze bespreking worden conclusies getrokken m.b.t. variabelen die in dit onderwijssysteem een belangrijke rol spelen. Tenslotte worden de vragenlijsten besproken (paragraaf 2.7).

2.2 Uitgangspunten.

Het evaluatiesysteem dat in dit proefschrift object van onderzoek is, voorziet docenten en andere belanghebbenden van onderwijskundige feedback teneinde de kwaliteit van het onderwijs te verbeteren. Het systeem vervult dus louter formatieve functies. De evaluatie van het onderwijsprogramma in de hier te bespreken aanpak verloopt in twee fasen. De eerste fase bestaat uit het doorlichten van cursussen op een aantal punten die van belang geacht worden bij het beantwoorden van de vraag in hoeverre het onderwijs in een bepaalde cursus voldaan heeft aan bepaalde criteria die verderop uitgebreid worden besproken. Aan het einde van iedere cursus vullen docenten en studenten een vragenlijst in die daarop betrekking heeft. De vragenlijst voor docenten heeft voor iedere cursus een identieke samenstelling. De vragenlijst voor studenten kan naast een aantal standaardvragen een aantal cursusspecifieke vragen bevatten. De docenten van de betreffende cursus dragen gewoonlijk deze cursusspecifieke

vragen aan. Binnen een periode van zes weken ontvangen de bij de cursus betrokkenen een evaluatierapport. In dit rapport worden de reacties op het onderwijs kort besproken en zijn gemiddelde scores op de vragenlijstitems opgenomen. De onderwijscommissie en de jaarcoördinatoren ontvangen eveneens dit evaluatierapport. Om vergelijkingen tussen cursussen mogelijk te maken, en om conclusies m.b.t. de relatieve kwaliteit van de cursus onder studie te vergemakkelijken zijn in het rapport grafieken op- genomen waarin voor alle cursussen de gemiddelde scores op de standaarditems worden weergegeven.
Op verzoek van de betrokkenen kunnen echter ook verdere analyses verricht worden om meer specifieke vragen over het functioneren van het onderwijsprogramma te beantwoorden. Dit is de tweede fase van de evaluatieprocedure. Als blijkt dat een programma kampt met structurele problemen, of dat een cursus uitermate slecht beoordeeld is, kan de onderwijscommissie van de medische faculteit besluiten om nader onderzoek te laten verrichten naar de oorzaak van de gesignaleerde problemen. Voor een dergelijk vervolgonderzoek wordt een keuze gedaan uit de methoden zoals besproken in paragraaf 1.3 te weten: het verrichten van observaties in de cursus, het interviewen van studenten en docenten, het analyseren van toetsen, of het afnemen van een op de problemen toegesneden enquête.

2.2.1 Het evaluatiesysteem in de context van de onderwijsorganisatie.

Het evaluatiesysteem van de medische faculteit heeft zoals gezegd de taak bestuurders en docenten van zodanige informatie voorzien dat de kwaliteit van het onderwijs verbeterd kan worden. Twee factoren spelen een rol bij de realisatie van deze doelstelling: de beschikbaarheid van betrouwbare en valide evaluatiegegevens en inzicht in het gebruik van die gegevens ten behoeve van de onderwijsorganisatie. In hoofdstuk 1 werd betoogd dat er in z'n algemeenheid een relatie bestaat tussen de structuur van de onderwijsorganisatie en de uitgangspunten van het evaluatiesysteem. In deze paragraaf zullen we een verband leggen tussen specifieke kenmerken van de onderwijsorganisatie van de medische faculteit en de opzet van het daarop afgestemde evaluatiesysteem.
De medische faculteit heeft, vanaf haar oprichting in 1974, een onderwijsprogramma ontwikkeld dat in vele opzichten afwijkt van andere medische curricula. In het programma wordt een grote nadruk gelegd op de zelfverantwoordelijkheid van studenten. Het programma heeft een interdisciplinaire opbouw en is vormgegeven volgens de principes van probleemgestuurd onderwijs (zie paragraaf 2.4 en verder).
De afwijkende vorm van het onderwijsprogramma heeft ertoe geleid, dat een organisatiestructuur is ontstaan waarin een aantal verantwoordelijkheden die traditioneel bij de individuele docent of vakgroep liggen, tot de facultaire verantwoordelijkheid zijn gaan behoren. De faculteitsraad, het

faculteitsbestuur en de beleidsadviescommissies (onderwijs- en examencommissies) hebben, meer dan bij andere faculteiten, een beslissende stem bij de voorbereiding en uitvoering van het onderwijsprogramma. In de nota 'Hoofdlijnen van de onderwijsorganisatie van de Faculteit der Geneeskunde' (BO 84.14200) wordt een uitgebreide beschrijving van deze organisatiestructuur gegeven. Voor de meeste rollen en functies van docenten, worden in deze nota nauwkeurige taakomschrijvingen gegeven.
Kenmerkend voor de onderwijsorganisatie van de medische faculteit is de projectstructuur en de centrale rol van de onderwijscommissie. Interdisciplinaire docententeams dragen in een aantal curriculumprojecten zorg voor de uitvoering van bepaalde taken die facultair vastgesteld zijn. Er bestaan twee projecten: het project 'Curriculum' en het project 'Curriculum Ondersteuning'. Het project 'Curriculum' bestaat uit vijf deelprojecten: het deelproject 'eerste studiejaar', het deelproject 'tweede studiejaar', het deelproject 'derde en vierde studiejaar', het deelproject 'vijfde en zesde studiejaar' en het deelproject 'keuze-onderwijs'. Het project 'Curriculum Ondersteuning' bestaat uit een aantal deelprojecten die de verdere ontwikkeling van het onderwijsprogramma ondersteunen, bijvoorbeeld het deelproject 'Training en Vorming' (docententraining), het deelproject 'Evaluatie van Studieresultaten', en het deelproject 'Programma-evaluatie'. De onderwijscommissie heeft het deelproject 'Programma-evaluatie' ingesteld om de docenten die deel uitmaken van het project 'Curriculum' van informatie te voorzien. De onderwijscommissie coördineert deze projecten.
In hoofdstuk 1 werd aan de hand van een artikel van Darling--Hammond, Wise en Pease (1983) een analyse gegeven van het verband tussen de structuur van onderwijsorganisaties en de rol van programma-evaluatie daarbij. Uit het bovenstaande moge blijken dat men de organisatiestructuur van de medische faculteit zou kunnen typeren als een 'rationalistisch model van een onderwijsorganisatie'. In een dergelijk organisatiemodel wordt onderwijs planmatig benaderd: docentenrollen zijn nauwkeurig beschreven, verantwoordelijkheden voor voorbereiding en uitvoering van het onderwijs liggen meer bij facultaire bestuursorganen dan bij individuele docenten. In de bovengenoemde nota (BO 84.14200) concludeert men dan ook dat de structuur van de onderwijsorganisatie van de medische faculteit de individuele docent slechts beperkte vrijheid laat. "Redenerend vanuit een willekeurig Maastrichts staflid, is het gevolg dat men vaker geacht wordt dingen te doen en te laten die door allerlei facultaire instanties zijn uitgedacht, terwijl de normale instelling van een staflid is in grote lijnen zijn inzichten en die van collega's in de directe werkomgeving te volgen.''
Het verplaatsen van een deel van de beslissingsbevoegdheid inzake structuur en inhoud van het onderwijsprogramma van individuele docenten naar facultaire bestuursorganen leidt ertoe dat de laatste behoefte krijgen aan gedetailleerde

informatie over het functioneren van dat programma. Programma-evaluatie vormt voor de facultaire bestuursorganen dus een instrument om de kwaliteit van het onderwijsprogramma te beheersen.

2.3 Instrumentontwikkeling.

Gegeven de keuze voor het gebruik van studentoordelen ten behoeve van verschillende evaluatie-doeleinden, is de vraag van belang hoe men met behulp van studentoordelen de kwaliteit van onderwijs kan meten: aan welke eisen dienen vragenlijsten te voldoen om op een betrouwbare en valide wijze de onderwijskwaliteit te meten, hoe worden de criteria geformuleerd om de kwaliteit van het onderwijs te bepalen en welke informatie moeten vragenlijsten opleveren voor docenten en bestuurders. De verschillende methoden om vragenlijsten op het gebied van onderwijsevaluatie te construeren hebben alle als uitgangspunt de vraag welke dimensies kenmerkend zijn voor kwalitatief goed onderwijs. Het evalueren van onderwijs impliceert namelijk dat men de geobserveerde onderwijspraktijk op een of andere manier beoordeeld aan de hand van een aantal criteria die aangeven hoe dat onderwijs er idealiter uit zou moeten zien. In de literatuur worden drie gangbare methoden genoemd voor het ontwikkelen van instrumenten die kenmerken van goed onderwijs meten (Jones, 1981; De Neve & Janssen, 1982; Marsh, 1982, 1984). De eerste methode baseert de constructie van de vragenlijst op algemeen gangbare theorieën of modellen over onderwijsleerprocessen. De tweede methode, die men als inductief zou kunnen beschrijven, berust op het inventariseren van opinies van bij het onderwijs betrokkenen, in casu docenten en studenten, over wat zij als kenmerken van goed onderwijs zien. Op basis van deze opinies worden vervolgens vragen geformuleerd. De laatste methode die in de literatuur beschreven wordt, is een combinatie van beide. In de eerste methode leidt men uit gangbare theorieën over onderwijs zoals die van Carroll (1963), Bloom (1976), Ausubel (1968), Gagne (1974) en Cooley en Lohnes (1976), af welke variabelen een rol spelen in het onderwijsleerproces, en welke beoordelingscriteria men dus zou moeten hanteren om de kwaliteit van het onderwijs te bepalen. Deze theorieën geven aanwijzingen voor het didactisch handelen van de docent, de vormgeving van de leerstof, de manier van toetsing. De theorie van Ausubel (1968) handelt bijvoorbeeld over de wijze waarop leerlingen nieuw betekenisvol materiaal integreren in het geheel van kennis dat zij reeds bezitten over het betreffende onderwerp. Hij formuleerde op basis van zijn theorie richtlijnen die betrekking hebben op de vormgeving van het onderwijsleerproces. De bekendste richtlijn is wel dat het onderwijs zoveel mogelijk moet aansluiten bij de al aanwezige cognitieve structuren van leerlingen. Als men zich bijvoorbeeld op de theorie van Ausubel zou baseren om de belangrijke variabelen binnen het onderwijsproces te meten, dan zou dat, in het geval dat men vragenlijsten zou gebruiken,

betekenen dat vooral vragen geformuleerd worden die betrekking hebben op de mate waarin in het gegeven onderwijs de aansluiting van de leerstof op de voorkennis van leerlingen bereikt wordt. Deze constructiewijze heeft echter weinig weerklank gevonden in de praktijk van de onderwijsevaluatie in het hoger onderwijs. De reden hiervoor moet gezocht worden in een gebrek aan consensus over de wijze waarop het onderwijs bijdraagt aan leren. Jones (1981) en Marsh (1984) beargumenteren dat theorieën over het onderwijs empirisch vaak gebrekkig onderbouwd zijn. Bovendien richten een aantal theorieën, met name de psychologische, zich op aspecten van leerprocessen die slecht vertaalbaar zijn naar de onderwijspraktijk. Tenslotte blijken de meeste theorieën zelden duidelijke omschrijvingen of definities van variabelen te geven. Jones (1981) geeft een voorbeeld van veel voorkomende kritiek op deze constructiewijze:
'Theoretical models of teaching exist, but such models are only as acceptable as the extent to which they are in line with popularly accepted criteria of "good" teaching. No-one would attach any credibility to a model for good teaching that rated Dr. X as competent when the students' examinations were poor, they were disenchanted with the course and when the other staff in Dr. X's department considered that his material was out-of-date. While this is an extreme example, it does suggest the two kinds of criteria against which it is possible to judge a model. These are: 1) the quantity and/or quality of students' learning; 2) the opinions of parties involved with the teaching/learning activity.' Het gebrek aan consensus over de vraag welke dimensies kenmerkend zijn voor goed onderwijs, heeft ertoe geleid dat men zijn toevlucht zocht in een andere aanpak. Op een inductieve wijze worden vragenlijsten geconstrueerd die items bevatten over kenmerken van goed onderwijs. Deze vragenlijsten bevatten gewoonlijk vragen die geconstrueerd zijn op basis van opinies van docenten en studenten over kenmerken van goed onderwijs. Teneinde dergelijke vragen te construeren, vraagt men veelal eerst aan studenten en docenten om een beschrijving te geven van goed docentgedrag, wat de kenmerken zijn van een goed opgezette cursus etc. Nadat de resulterende lijst een aantal malen is afgenomen, worden met behulp van factoranalyses de onderliggende dimensies in de vragenlijst onderscheiden. Deze dimensies worden gerepresenteerd door vragen die een bepaald gezamenlijk kenmerk van het onderwijs meten (bv de manier van leerstofpresentatie, de kwaliteit van de gebruikte leermaterialen). Kulik en McKeachie (1975) concludeerden in een overzichtsartikel dat de volgende zes factoren meestal in de vragenlijsten gevonden worden:

1. 'skill', bekwaamheid/doceervaardigheid van een docent,
2. 'structure', organisatie en voorbereiding van cursus of colleges,
3. 'workload', zwaarte van de cursus, eisen die de docent stelt,
4. 'rapport', relatie van de docent tot individuele

student, vriendelijkheid, aandacht,

5. 'instructor-group interaction', mate waarin de docent erin slaagt een atmosfeer te scheppen waarin studenten hun meningen durven te ventileren, zowel tegenover de docent als tegenover elkaar,
6. 'feedback', de door de docent verstrekte informatie over de kwaliteit van door studenten gegeven antwoorden of het geleverd werk.

Marsh (1984) daarentegen noemt, in een overzichtsartikel over een veelvuldig gebruikte vragenlijst voor cursusevaluatie, negen dimensies. Aan bovengenoemde factoren voegt hij de volgende drie toe:

7. 'learning/ value', waardering die studenten hebben voor het vak,
8. 'enthusiasm', het enthousiasme van de docent,
9. 'breadth of coverage', mate van diepgang van de behandelde leerstof.

Hoewel de factoranalytische methode veelvuldig gebruikt wordt teneinde op empirische wijze tot een beschrijving van kenmerken van onderwijs te komen, kleven er een aantal bezwaren aan (Larson, 1979; Van Rooijen & Vlaander, 1983; Whitely & Doyle, 1976; Cohen, 1981; Dowell & Neal, 1982; De Neve & Janssen, 1982). Deze bezwaren hangen nauw samen met de vraag in hoeverre men m.b.v. factoranalyse feitelijke dimensies binnen onderwijsgedrag kan onderscheiden. De gevonden factoren kunnen namelijk ook een weerspiegeling zijn van de impliciete theorieën die beoordelaars hanteren over onderwijsgedrag. Studenten kunnen hun eigen set van theorieën of assumpties hebben over de manier waarop onderwijs hoort te functioneren. Als men studenten vraagt een docent te beoordelen kunnen hun antwoorden meer beïnvloed zijn door deze impliciete theorieën, en minder berusten op feitelijk onderwijsgedrag van de docent. Harari en Zedeck (1973) zijn van mening dat beoordelingsvragenlijsten die op deze wijze zijn geconstrueerd, meestal vage beschrijvingen van docentengedrag bevatten en over het geheel genomen te globale informatie geven. De discussie in hoeverre studenten als experts beschouwd kunnen worden bij de beoordeling van onderwijsgedrag, heeft er onder andere toe geleid dat men een combinatie van beide aanpakken is gaan voorstaan (De Neve & Janssen, 1982; Marsh, 1984). Een gecombineerde aanpak bestaat daarin dat in de vragenlijst van de ene kant variabelen opgenomen worden die in theorieën m.b.t het onderwijsleerproces beschreven worden, terwijl van de andere kant ook op empirische wijze verkregen kenmerken, die door de onderzochte groep als belangrijk worden beschouwd, in vragen vertaald worden. Een dergelijke aanpak wordt beschreven door De Neve en Janssen (1982). Zij ontwikkelden een vragenlijst die enerzijds gefundeerd is in het model van didactische analyse van Van Gelder (1975) en anderzijds op studentopinies over onderwijs. Het model van Van Gelder bevat variabelen die betrekking hebben op aspecten van het onderwijs, zoals beginkenmerken van studenten, de in-

structiemethode en de evaluatie. In de resulterende vragenlijst die zich concentreerde op doceergedrag werden een aantal factoren onderscheiden die studenten hanteerden bij de beoordeling van docenten. Tenslotte werd onderzocht in hoeverre de factoren die in de studentenvragenlijst naar voren kwamen, pasten in het model van Van Gelder. Op deze manier kon men een vragenlijst construeren die paste binnen dit model en rekening hield met percepties van studenten over doceergedrag. Een laatste punt waarmee rekening gehouden moet worden bij de constructie van een vragenlijst is het soort informatie dat verkregen wordt. Volgens Rotem en Glasman (1979) is het door onderzoekers algemeen geaccepteerd dat beschrijvende items de voorkeur verdienen boven evaluerende items. De eerste zijn minder bedreigend voor docenten dan de laatste. Zij noemen verder nog vier eisen waaraan beoordelingsvragenlijsten moeten voldoen, om informatie te leveren die bruikbaar is voor docenten en bestuurders:

1) in de vragenlijst moeten items opgenomen zijn die aspecten van het onderwijs weergeven waarop docenten invloed kunnen uitoefenen,
2) de vragenlijst moet items bevatten die beschrijving geven van specifiek gedrag en niet van algemene persoonlijkheidstrekken die moeilijk te veranderen zijn,
3) de items moeten zo concreet mogelijk zijn,
4) de items moeten relevant zijn voor de betreffende cursus.

2.4 Beknopte historie van probleemgestuurd onderwijs.

De oudste methode waarbij groepjes leerlingen aan de hand van problemen tot studie gestimuleerd werden, is het eerst beschreven door de Griekse schrijvers Xenophon en Plato. Zij beschrijven hoe de uit Athene afkomstige wijsgeer, Socrates, zijn tijd doorbracht met disputeren met omstanders, en jongeren in de filosofie onderrichtte (Russel, 1946).
Socrates gebruikte als voornaamste onderwijsmethode het principe van de dialectiek: een methode om kennis te verwerven door middel van vraag en antwoord. Hij verzamelde een groepje leerlingen om zich heen en gaf aan de hand van problemen die hij hun voorlegde onderwijs. Met behulp van deze aanpak - Socratische methode genoemd- lukte het hem willekeurige omstanders de stelling van Pythagoras te laten bewijzen, door hen vragen te stellen en hen vervolgens met de konsekwenties van hun antwoord te confronteren. Volgens Plato was deze methode bruikbaar voor het leren doorzien van problemen die aan twee condities voldoen: het besprokene moet meer een logisch dan een feitelijk karakter dragen en de leerlingen moeten tevens over voldoende kennis beschikken met betrekking tot het probleem.
De filosofie die aan probleemgestuurd onderwijs ten grondslag ligt, kan voor een deel teruggevonden worden in Dewey's opvattingen over onderwijs (Dewey, 1916). Hij was een voorstander van het gebruik van ''real life situations'' in het

onderwijs. Studenten moesten problemen uit de praktijk leren analyseren en oplossen. Dat verdiende de voorkeur boven het leren van losse feiten. Leren was niet in de eerste plaats een hoeveelheid kennis verwerven, maar het zich eigen maken van oplossingsmethoden. Dewey (1916) maakte gebruik van ''projecten'' om kinderen op school te laten leren hoe uitvindingen tot stand waren gekomen, en hoe de cultuur zich ontwikkelt.
Katona (1940) verrichtte onderzoek naar de wijze waarop in het traditionele onderwijs informatie overgedragen wordt. Hij concludeerde op basis van zijn onderzoeksresultaten dat het onderwijs teveel 'ready made solutions' aanbood. Hij toonde aan dat studenten veel kennis verwerven die voor hun in belangrijke mate betekenisloos is, hetgeen ertoe leidt dat ze die kennis sneller vergeten. Katona was dan ook van mening dat het onderwijs kennis vaak op ineffectieve wijze overdraagt (Glaser, 1984).
Dit probleem werd ook onderkend aan de Harvard universiteit en de Harvard Business School. Fraser (1931) beschrijft dat docenten die verbonden waren aan deze instellingen, tot de conclusie kwamen dat studenten slechts marginaal in staat waren kennis die zij tijdens hun studie hadden opgedaan, toe te passen. Men was daarom van mening dat een curriculum leersituaties moest verschaffen die studenten in staat stelden kennis te leren hanteren (Schmidt, 1982). Er werd een methode ontwikkeld, de 'case-study method', waarbij aan studenten een probleembeschrijving (case) werd voorgelegd die zij moesten analyseren. Vervolgens werd van hen gevraagd relevante informatie te verzamelen en oplossingen voor het geval te formuleren.
De aandacht voor de vraag hoe men bij studenten op een zodanige wijze kennis kan overbrengen, zodat deze voor hun 'bruikbaar' wordt, leidde ertoe dat ook andere methoden ontwikkeld werden die aan dit probleem tegemoet trachtten te komen. Bruner (1961) ontwikkelde een voorloper van probleemgestuurd leren waarbij hij leerlingen confronteerde met problemen die zij moesten oplossen door middel van discussie met medeleerlingen. Deze methode van ontdekkend leren, (''learning by discovery''), zou tot een dieper inzicht in de werkelijkheid leiden, gelegenheid geven tot het oefenen van cognitieve vaardigheden, de intrinsieke motivatie verhogen en het verwerken en onthouden van kennis bevorderen (Schmidt, 1982).
De eerste probleemgestuurde medische cursus werd in 1961 geconstrueerd aan de Case Western University, Cleveland, Ohio. Het betrof een haematologie cursus. In 1962 werd een evaluatiestudie gepubliceerd waarin vergelijkingen werden gemaakt tussen studenten die de probleemgestuurde cursus hadden gevolgd en studenten die de haemotologiecursus in traditionele vorm hadden gevolgd. De conclusie van dat onderzoek was dat studenten die de probleemgestuurde cursus hadden gevolgd, op een kennistoets beter presteerden dan de andere groep studenten (Harris, Horrigan, Ginther & Ham, 1962).
Het eerste volledig probleemgestuurde medisch curriculum werd

in 1965 voorbereid aan de McMaster University in Hamilton, Canada. In 1969 begonnen de eerste studenten met hun studie (Neufeld & Barrows, 1974; Sweeney et al, 1975; Hamilton, 1976; Barrows & Tamblyn, 1980).
Uitgangspunten van het programma waren, en zijn, de nadruk op probleemoriëntatie, op zelfwerkzaamheid van studenten, en op integratie van de leerstof uit verschillende medische disciplines. In het programma wordt veel aandacht besteed aan het leren gebruiken van medische kennis. Niet alleen het verwerven van feitenkennis is belangrijk, maar ook het gebruik van die kennis in probleemsituaties. Barrows (in Barrows & Tamblyn, 1980) beschrijft in het onderstaande citaat een ervaring, die illustratief is voor het denken over het medisch onderwijs in de jaren zestig, en de daaruit voortvloeiende ideeën over veranderingen in het medisch onderwijs. ''In 1963, I had been responsible for several years for a neurological clinical clerkship through which six or more third year medical students percolated every four weeks. I became concerned that the usual faculty evaluations were not providing data that were truly helpful to the student. As a result, the simulated patient was developed and used as a standardized patient problem; this provided more data concerning student competence. It revealed that, although students had, for the most part, good techniques in performing a neurological history and physical examination, they seemed to have a paucity of basic knowledge that they could apply to the patient problem. This seemed paradoxical to me, as I had been closely associated with and contributed to, the students' prior courses in neuroanatomy, neurophysiology, and clinical neurology. I knew that these students had been exposed to, and had passed, excellent, detailed courses.
This observation about students was shared by many on the faculty, leading to the recurrent, half-serious suggestion that the school ought to have an "inverted curriculum" where the students would have two years of patient exposure and then two years of basic science. Students thus could enhance their learning and application of information, since the importance and relevance of basic science information could be perceved more readily.''
De medisch faculteit te Maastricht, samen met de medische faculteit te Beersheva in Israël, was in 1974 de tweede faculteit die haar onderwijsprogramma volgens de principes van probleemgestuurd onderwijs vorm gaf. Het onderwijsprogramma van McMaster heeft in een aantal opzichten grote invloed gehad op de vormgeving van het Maastrichtse programma: de nadruk op zelfwerkzaamheid van studenten, op probleemoriëntatie, het leggen van een zwaartepunt bij de eerstelijns gezondheidszorg, de aandacht voor attitude-ontwikkeling, een vroegtijdige intensieve deelname van studenten aan de gezondheidszorg, streven naar integratie van verschillende disciplines, op de vormgeving van onderwijsleersituaties en op opvattingen over de rol van docenten. In de volgende paragraaf zal het programma van de medische faculteit te Maas-

tricht met haar nadruk op deze elementen beschreven worden.

2.5 Het probleemgestuurde medisch curriculum van de Rijksuniversiteit Limburg.

Het onderwijsprogramma van de medische faculteit omvat een tijdspanne van zes jaar. De eerste vier studiejaren zijn opgebouwd uit 20 verplichte. Elk blok duurt zes weken en is opgezet rondom een centraal thema. Tevens maken vier keuzeblokken van zes weken deel uit van dit programma. De laatste twee jaren zijn gereserveerd voor stages (o.a. co-assistentschappen) en keuze-onderwijs.
In het programma staan niet de verschillende medische vakgebieden maar medische probleemgebieden of thema's centraal. Vandaar dat men in het onderwijsprogramma tevergeefs zal zoeken naar vakgebieden als fysiologie, anatomie, biochemie, chirurgie, interne geneeskunde, medische psychologie etc. De basis- en de klinische vakken komen aan de orde in het kader van de gekozen blokthema's. Aan de hand van een probleem (bijvoorbeeld hoofdpijn) bestuderen studenten relevante vakgebieden als neuro-anatomie en neurofysiologie. Het derdejaars blok (blok 3.1; academiejaar 1984/1985) met als thema koorts, infecties en ontstekingen is bijvoorbeeld opgebouwd rond de volgende subthema's:

week 1 en 2: koorts met klachten van de bovenste luchtweg,
week 3: koorts met buikklachten,
week 4: koorts en aandoeningen van de urinewegen,
week 5: genitale infekties,
week 6: koorts met diverse klachten.

In dit blok ligt niet de nadruk op het klinische beeld van een aantal ziekten, maar op de pathofysiologische mechanismen, die aan die ziekten ten grondslag liggen. Binnen een subthema worden studenten geconfronteerd met relevante problematiek. In het genoemde blok 3.1 bijvoorbeeld hebben de aangeboden taken in week 1 en 2 betrekking op problematiek rond neus, keel, oor en in mindere mate op de longen. Van de student wordt verwacht dat hij zich verdiept in de anatomie en fysiologie van de bovenste luchtwegen en dat hij therapeutische mogelijkheden bestudeert om verschillende bacteriële en virale aandoeningen te bestrijden.
De blokthema's zijn weer gerelateerd aan centrale jaarthema's. In het eerste jaar is dit het thema 'orientatie en inleiding in de geneeskunde'. In het tweede jaar staat het normaal functioneren van de mens in verschillende levensfasen centraal. De blokken uit het derde en vierde jaar zijn opgebouwd rond thema's die de meest voorkomende en belangrijke klachten of verzamelingen klachten uit de eerstelijns gezondheidszorg weerspiegelen. In het vijfde en zesde jaar ligt zoals gezegd de nadruk op practische stages, zoals de psychomedische stage (pmo-pms), de stage huisartsgeneeskunde (pmo-h) en de klinische stages (pmo-k) interne geneeskunde,

chirurgie, gynaecologie en obstetrie, paediatrie, kno, oogheelkunde, neurologie en dermatologie. Tevens besteden studenten een gedeelte van het vijfde en zesde jaar aan keuze-onderwijs (totale tijdsduur 14 weken).
Het theoretische onderwijs aan de medische faculteit wordt gekenmerkt door een grote nadruk op probleemorientatie (d.w.z. dat problemen uit de medische praktijk centraal gesteld worden) en zelfwerkzaamheid van studenten. De nadruk op probleemorientatie komt vooral naar voren in de keuze van de blokthema's en de inhoud van de blokken. Zelfwerkzaamheid van studenten wordt bevorderd doordat studenten zelf keuzen kunnen maken t.a.v. de te bestuderen leerstof, en doordat het aantal contacturen zeer beperkt is.
De gehanteerde onderwijsmethode stoelt op de principes van probleemgestuurd onderwijs (Barrows & Tamblyn, 1980; Schmidt, 1982; Schmidt & De Volder, 1984). Studenten werken in kleine groepen (bestaande uit 8 a 10 personen), onder begeleiding van een staflid, aan problemen (ook wel taken genoemd) deels afkomstig uit de medische praktijk. In eerste instantie probeert de groep op basis van aanwezige voorkennis een analyse van het probleem te maken. Tijdens de analyse zullen vragen opkomen waarvan de beantwoording om nadere studie vraagt. De opmerkingen en vragen die tijdens de discussie naar voren komen, vormen de basis voor het formuleren van leerdoelen. Deze leerdoelen leiden tot het verzamelen van informatie die relevant is ten aanzien van het probleem. Na de fase van informatieverzameling worden in de volgende groepsbijeenkomst de resultaten uitgewisseld en besproken.
De taken zijn gebundeld in een zogenaamd blokboek.

2.5.1 Het blokboek.

Het blokboek wordt in opdracht van de onderwijscommissie samengesteld door een planningsgroep. Een planningsgroep is een multidisciplinair team van docenten, aangevuld met studenten. Ze bestaat uit maximaal drie docenten en twee studenten. Het blokboek kan als een werkboek voor studenten beschouwd worden. De kern van het blokboek wordt gevormd door een aantal verschillende problemen of taken die voortvloeien uit het blokthema. Het blokboek bevat tevens literatuursuggesties, toelichtingen op het thema, een lijst met inhoudsdeskundigen (= stafleden met specialistische kennis m.b.t. een bepaald onderwerp), een lijst met beschikbare Audio-Visuele-middelen, het programma van het skillslab, en van het vaardighedenlaboratorium, roosters, en een indeling in onderwijsgroepen. In tabel 2.1 is als voorbeeld een inhoudsopgave uit blokboek 3.1 (1984/1985) weergegeven.

Tabel 2.1: Inhoudsopgave blokboek 3.1., academiejaar 1984/1985.

2.5.2 Taken.

Taken kunnen beschouwd worden als opdrachten die aan onderwijsgroepen gegeven worden met de bedoeling bepaalde, door de planningsgroep gewenste, studie-aktiviteiten te stimuleren. Er zijn een aantal pogingen gedaan de in het programma gebruikte taken te categoriseren. Schmidt en Bouhuijs (1980) maken een onderscheid naar de leerstof waarop taken betrekking hebben, naar de aard van de denkprocessen die uitgelokt worden, naar het soort verschijnselen die feitelijk beschreven worden, en naar de activiteiten die van studenten gevraagd worden. Zij kiezen voor een indeling die gebaseerd is op de activiteiten die van studenten gevraagd worden. Ze komen dan tot het volgende onderscheid: de probleemtaak, de toepassingstaak, de actietaak, de discussietaak, en de studietaak. De probleemtaak bestaat volgens Schmidt en Bouhuijs (1980) uit: 'een min of meer neutrale beschrijving van een aantal verschijnselen of gebeurtenissen die in een zekere relatie met elkaar lijken te staan'. De taak van de student is deze verschijnselen in onderlinge samenhang te verklaren.In de toepassingstaak wordt van studenten gevraagd eerder verworven kennis toe te passen: deze taakvorm wordt soms gebruikt voor het behandelen van statistische- of epidemiologische problemen. Als er van studenten gevraagd wordt handelend op te treden, is er sprake van een aktietaak (bijvoorbeeld: het rollenspel, of aktiviteiten in de praktijk van de gezondheidszorg). Bij de discussietaak staat het gezamelijk bediscussiëren van kennis centraal (te denken valt aan medisch-ethische vraagstukken). De studietaak kan beschouwd worden als een studie-opdracht: 'bestudeer de volgende hoofdstukken uit het boek van ..'. Zie voor een meer uitgebreide

bespreking van deze taakvormen Schmidt en Bouhuijs (1980). Snellen-Balendong (1984) presenteert een indeling gebaseerd op vormkenmerken van de taak. Zij onderscheidt onder andere de volgende vormen: het artikel of citaat, het fenomeen, de opdracht, POMR (problem-oriënted medical record), probleempakket, relaas, SOEP-casus.
Veel gebruikte vormen van taken zijn, in studiejaar 1 en 2, het relaas en het fenomeen. In studiejaar 3 en 4 komt vooral de SOEP-casus, de POMR en het probleempakket voor. In onderstaande tabellen is ter illustratie een relaas en een SOEP-casus weergegeven. Een relaas is een uitgebreide beschrijving van een situatie of gebeurtenis in de ruimste zin, echter meestal met een patient als hoofdpersoon. Een SOEP-casus bevat een korte casuïstische probleembeschrijving waarna de Subjektieve en Objektieve gegevens van de patient gerangschikt moeten worden teneinde een Evaluatie en Plan op te stellen.

Tabel 2.2: Voorbeeld van een relaas uit blok 2.2 (Casus Annemarie), academiejaar 1984/1985

Annemarie is het dochtertje van een Nederlands echtpaar dat over de hele wereld heeft gezworven. Zij werd onlangs 7 jaar en omdat haar ouders zich definitief in Nederland vestigden, werd ze aangemeld bij de lagere school. De eerste maanden viel het de juffrouw op dat Annemarie houterig beweegt en moeite heeft met het goed hanteren van potlood en schaar. Rekenen gaat haar vlot af, lezen heel hakkelend, net als haar spontane spraak. In overleg met de ouders wordt Annemarie gezien door de schoolarts. De schoolarts onderzoekt Annemarie en vraagt zich af of ze spastisch is. De moeder vertelt dat ze tijdens de zwangerschap van Annemarie veel last van hoge bloeddruk heeft gehad en veel moest rusten. Omdat de geboorte niet vorderde en het kind in nood geraakte, moest er met spoed een keizersnede verricht worden. Toen Annemarie daarop geboren werd, huilde ze niet direct door en moest ze kortdurend met masker en ballon beademd worden. Toen ze 22 maanden was, ging ze zelf lopen. Vaak struikelt ze daarbij over haar eigen benen en hollen heeft ze nooit goed gekund. Omdat de ouders vele standplaatsen bekleedden, sprak ze met woorden van vele lokale talen doorspekt met wat Nederlands en Engels. Schoolbezoeken duurden hooguit een half jaar voordat er weer verhuisd moest worden. Na het gesprek met de schoolarts wordt in overleg met de huisarts besloten om Annemarie verder te laten onderzoeken.

Tabel 2.3: Voorbeeld van een SOEP-casus uit blok 4.2, academiejaar 1984/1985

Taak 1: inleiding
Bij deze casus gaat het niet zozeer om het vinden van een diagnose, maar om de differentiaaldiagnose van dit bovenbuiksbeeld.

CASUS

Je bent huisarts en je hebt avonddienst.
Je wordt om 22.00 uur gebeld door de vrouw van een patiënt van een van je collegae. Ze vertelt dat haar man die middag buikpijn gekregen heeft. Hij heeft dit enkele uren afgewacht maar de pijn wordt steeds erger en ze wil graag dat je komt kijken.
Bij aankomst blijkt de man in bed te liggen met zijn benen opgetrokken. Hij ligt het liefst stil want bewegen doet hem pijn, vooral in de bovenbuik.

1. *Welke anamnestische gegevens zou je willen weten?*
2. *Maak hier een overzichtelijke lijst van.*
3. *N.B.: De tutor beschikt over aanvullende gegevens.*
4. *Pas daarna doorgaan op volgende bladzijde.*

Lichamelijk onderzoek.

Je vindt een zieke man met veel pijn in de bovenbuik. Bij onderzoek van het abdomen zijn er duidelijke tekenen van 'peritoneale prikkeling' diffuus in de bovenbuik. Hierdoor is diepe palpatie niet goed mogelijk. Peristaltiek is schaars aanwezig.

Temperatuur	*38° c*
Bloeddruk	*110/80 P.: 120*
Onderzoek hart/longen	*geen afwijkingen*

Analyseer alle gegevens op schematische wijze (m.b.v. het S.O.E.P. schema). De nadruk dient hierbij te liggen op:
a) een uitgebreide differentiaal diagnose van dit beeld;
b) het verder door de huisarts te volgen beleid.

Vragen:
1. Waarop wijst peritoneale prikkeling?
2. Welke tekenen van peritoneale prikkeling ken je?

De vorm van de taken of problemen kan bepalend zijn voor de manier waarop onderwijsgroepen aan de taak werken en welke leerdoelen zij formuleren (Gijselaers & Schmidt, 1985a, 1985b; Snellen-Balendong, 1985; Neame, 1984).

Uit onderzoek van Gijselaers en Schmidt (1985a) naar de relatie tussen het functioneren van een blok en de aard van de taken (wat betreft de vormgeving) bleek, bij ongewijzigde blokdoelstellingen, dat verandering van taakvorm studenten meer motiveerde tot studeren, en onderwijsgroepen meer tot discussie gestimuleerd werden. Dit had als gevolg dat studenten en docenten het blok positiever beoordeelden.
Onderzoek naar de aard van de relaties van belangrijke variabelen binnen probleemgestuurd onderwijs (aard van de taken, onderwijsgroep, tutor) laat zien dat de vorm van de taken een grote invloed heeft op het groepsfunctioneren (Gijselaers & Schmidt, 1985b). Een van de conclusies was dat onderwijsgroepen op een meer gestructureerde wijze werken naarmate taken beter geconstrueerd zijn, (bijvoorbeeld omdat rekening is gehouden met de voorkennis van studenten) duidelijke relaties met andere taken hebben, stimuleren tot zinnige groepsdiscussies en leiden tot concrete leerdoelen.
Schmidt en Bouhuijs (1977) onderzochten hoe de structurering van taken effect had op het leerresultaat en de satisfactie van studenten. Een groep studenten werkte aan taken zonder toegevoegde vragen. Dit werd beschouwd als de ongestructureerde conditie, een andere groep kreeg problemen met toegevoegde vragen (gestructureerde conditie). De groepen bleken niet te verschillen in tevredenheid met de eigen werkwijze en die van de onderwijsgroep. Studenten uit de ongestructureerde conditie bleken echter beter in staat te beoordelen in hoeverre begrippen uit een speciaal voor dat doel geconstrueerde begrippentoets samenhingen met de bestudeerde problemen. Op een later afgenomen retentietoets waren geen verschillen meer constateerbaar.
Neame (1984) benadrukt dat taken ingebed moeten zijn in een frame met andere taken, die tevens niet te snel tot oplossingen, zoals bijvoorbeeld een diagnose, mogen leiden. Volgens Neame bestaat namelijk bij studenten de neiging om de diagnosestelling als doel van een taak te zien.
Schmidt (1979) beschrijft een aantal criteria waaraan taken moeten voldoen. Taken mogen qua structuur niet te complex zijn (teveel informatie bevatten), ze moeten aansluiten op de voorkennis van studenten, stimuleren tot een zinnige groepsdiscussie, leiden tot de formulering van zo concreet mogelijke leerdoelen en tot zelfstudie-aktiviteiten.

2.5.3 Onderwijsgroep

Gedurende zes weken (de tijdsduur van een blok) komen de studenten twee maal twee uur per week in onderwijsgroepen van 8 à 10 studenten bijeen om aan taken van het hierboven beschreven type te werken. Deze groepen worden iedere zes weken opnieuw aselect samengesteld, dus niet op basis van individuele voorkeuren. Dit heeft als voordelen dat studenten in de loop van hun studie veel medestudenten leren kennen, en met collega's leren samenwerken die verschillende of uiteenlopende kennis, vaardigheden en opvattingen hebben. Bovendien

blijven de gevolgen van een slecht klikkende samenwerking beperkt tot een overzienbare periode.
Studenten dragen in belangrijke mate de verantwoordelijkheid voor de voortgang van de werkzaamheden in de onderwijsgroep. Om de voortgang van de onderwijsgroep zo optimaal mogelijk te laten zijn, verwacht men dat studenten op een systematische wijze de taken behandelen, duidelijke afspraken maken omtrent de te bestuderen stof, en een aktieve bijdrage leveren aan de groepsdiscussie. Bij toerbeurt fungeert een student als gespreksleider. In de onderwijsgroep worden de taken uit het blokboek geanalyseerd, afspraken gemaakt over de te bestuderen onderwerpen, en wordt de bestudeerde leerstof nabesproken.
Bouhuijs, Gijselaers en Kerkhofs (1984) onderzochten in hoeverre het slecht functioneren van onderwijsgroepen te wijten zou zijn aan de samenstelling van onderwijsgroepen. De vraag was of bepaalde studenten een permanent slechte invloed hebben op onderwijsgroepen, in die zin dat zij aanleiding geven tot groepsconflicten, geen positieve bijdrage leveren aan groepsdiscussies, afspraken slecht of niet nakomen, etc. Uit dit onderzoek bleek dat er geen reden bestond om aan te nemen dat sommige studenten een consistent negatieve invloed hadden op een onderwijsgroep. Met andere woorden het dysfunctioneren van deze groepen wordt niet bepaald door steeds dezelfde studenten.

2.5.4 De tutor.

De begeleider van een onderwijsgroep wordt tutor genoemd. Zijn handelen is erop gericht het leerproces van de groep op alle mogelijke manieren te bevorderen. Hij vervult daarbij noch de rol van gespreksleider, noch die van de klassieke docent die uitlegt hoe de leerstof in elkaar zit, die vragen en de studievoortgang controleert. De rol van de tutor kan beter beschreven worden met de termen groepsbegeleider of 'learning facilitator' (Rudduck, 1978).
De tutor heeft verschillende taken in een onderwijsgroep (Moust & Schmidt, 1985; Schmidt & Bouhuijs, 1980). De medische faculteit heeft deze taken tamelijk ondubbelzinnig in een nota omschreven (B.O. 84.14200). In tabel 2.4 zijn die functies samengevat.

Tabel 2.4: Taken van de tutor.

De tutor heeft:
- kennis van en inzicht in probleemgestuurd leren;
- vaardigheid en kennis met betrekking tot het functioneren van een groep.
- kennis van de opbouw en van de bedoeling van het betreffende blok en (in iets mindere mate) van het betreffende curriculumjaar.

Van de tutor worden de volgende aktiviteiten verwacht:
1. De tutor speelt een actieve en stimulerende rol binnen de onderwijsgroep;
- hij dringt aan op het maken van duidelijke afspraken;
- hij noteert de gemaakte afspraken en controleert of deze zijn nagekomen;
- hij evalueert de werkwijze van de groep en komt met eventuele voorstellen tot verbetering;
- hij stimuleert de probleem-analyse door de groep;
- hij zorgt voor een goede en efficiënte gespreksleiding door een van de groepsleden;
- hij stimuleert regelmatige samenvattingen;
- hij let op efficiënt gebruik van de tijd die voor de bijeenkomst van de onderwijsgroep beschikbaar is.

2. De tutor ziet toe op de voortgang in de onderwijsgroep;
- hij stimuleert de deelnemers van de groep die niet voldoende actief deelnemen;
- hij probeert de groep ertoe te brengen controle uit te oefenen op de werkzaamheden van individuen;
- hij ziet erop toe dat iedereen altijd aanwezig is en gaat na waarom iemand afwezig is.

3. De tutor heeft voldoende kennis van de opbouw en de bedoeling van het blok;
- hij is op de hoogte van de doelstellingen van het blok;
- hij is op de hoogte van de verschillende bronnen welke de planningsgroep beschikbaar heeft (gemaakt) voor het bereiken van de gestelde doelstellingen;
- hij moedigt het doeltreffend raadplegen van inhoudsdeskundigen aan;
- hij legt (eventueel) zelf contact met inhoudsdeskundigen;
- hij levert de planningsgroep snel feedback over de inhoudelijke voortgang van de onderwijsgroep;
- hij stimuleert het gebruik van zelfevaluatie, zowel individueel als in de groep;
- hij probeert te verhinderen dat de groep zich tevreden stelt met oppervlakkige probleem-analyse;

4. De tutor fungeert als eerste contactpersoon tussen de faculteit en de studenten.
- hij verzorgt organisatorische extra activiteiten van de onderwijsgroep (bespreking AV-ruimte, ruimte voor extra bijeenkomst van de groep)
- hij administreert de presentie van de studenten;
- hij tracht student-problemen op te lossen;
- hij signaleert wanneer een student zich binnen de groep problematisch gedraagt; probeert eventuele probleem-studenten weer tot de groep te brengen (eventueel) in overleg met de planningsgroep en/of jaarcoördinator;
- hij vervult mede een taak bij het afnemen van de bloktoets.

In principe kunnen alle wetenschappelijke medewerkers van de medische faculteit, arts of geen arts, als tutor fungeren. Tutoren worden speciaal voor hun taak getraind in een zogenaamde tutortraining (Schmidt & Moust, 1983). In deze training leren zij het onderwijssysteem kennen en worden zij getraind in een aantal vaardigheden die noodzakelijk zijn om de tutorrol te vervullen (groepsconflicten behandelen, technieken van vragenstellen, etc).
De vraag in hoeverre een tutor deskundig moet zijn op het gebied van de leerstof die in het betreffende blok aan de orde is, is een regelmatig terugkerend discussiepunt. Het gaat hier in feite om de vraag of een tutor die geen of onvoldoende kennis van de inhoud van een blok heeft, in staat is een groep adequaat te begeleiden. Daarmee bedoelt men of hij de groepsdiscussie kan bewaken wat betreft relevantie, inhoudelijke correctheid en diepgang. Onderzoek van De Volder en Schmidt (1981) naar de vraag in hoeverre de inhoudelijke deskundigheid van de tutor bijdraagt tot een optimale bewaking van het groepsproces, toonde aan dat er een positieve relatie is tussen de mate waarin tutoren zichzelf inhoudsdeskundig achten ten aanzien van het blok waarin zij als tutor optreden, en het oordeel dat studenten over die tutoren hebben. Tutoren die zich min of meer deskundig achtten, ontvingen van studenten positievere oordelen dan tutoren die zich weinig deskundig achtten. Op het functioneren van onderwijsgroepen en de daaruit resulterende studieprestaties van de groepen bleek de mate van inhoudsdeskundigheid echter nauwelijks invloed te hebben.
Tutoren die gebruik maken van hun deskundigheid om vragen te stellen in de onderwijsgroep of om rechtstreeks informatie over te dragen, kunnen in een situatie terecht komen waarin hun bijdrage die van studenten overheerst (Moust, De Grave & Gijselaers, 1985).

2.5.5 Werkwijze van de onderwijsgroep.

De gebruikelijke gang van zaken bij het behandelen van een taak is dat studenten eerst discussiëren over de taak, dan leerdoelen formuleren en vervolgens in de tijd tussen de onderwijsgroepsbijeenkomsten gaan studeren aan de hand van de leerdoelen. Men verwacht van onderwijsgroepen dat zij op een systematische manier een taak aanpakken (met behulp van de Zevensprong of het SOEP-systeem). De Zevensprong is een aanpak die geschikt is om taken aan te pakken waarin patientenproblemen of (bio)medische fenomenen gepresenteerd worden. Deze aanpak, die uit zeven stappen bestaat is in onderstaande tabel beschreven.

Tabel 2.5: De Zevensprong (Schmidt & Bouhuijs, 1980).

Stap 1: Helder onduidelijke termen en begrippen op.
Stap 2: Definieer het probleem.
Stap 3: Analyseer het probleem.
Stap 4: Inventariseer op systematische wijze de verschillende verklaringen die uit stap 3 naar voren zijn gekomen.
Stap 5: Formuleer leerdoelen.
Stap 6: Zoek aanvullende informatie buiten de groep.
Stap 7: Synthetiseer en test de nieuwe informatie.

De Zevensprong is een methode die voornamelijk de eerste twee studiejaren wordt toegepast. In het derde en vierde jaar gebruiken studenten vaak het SOEP-systeem. Deze aanpak wordt gebruikt bij taken waarin patientenproblemen beschreven worden en men van studenten verwacht dat zij uiteindelijk een diagnose en behandelingsplan opstellen (zie bijvoorbeeld de in pararagraaf 2.5.2 beschreven taak uit blok 4.2). Van de studenten wordt verwacht dat ze gezamenlijk bedenken welke vragen er aan de patient/familie gesteld moeten worden: dus het verzamelen van Subjektieve gegevens (de S van SOEP). Daarna moeten studenten bedenken welk Onderzoek zij willen verrichten (het verzamelen van objektieve gegevens), vervolgens moet het totaal aan beschikbare gegevens geëvalueerd worden waarna een plan ter behandeling wordt opgesteld. Al deze activiteiten moeten leiden tot zelfstudie die voor het diepgaand begrijpen van de casus noodzakelijk is.

2.5.6 Oriëntatie op de praktijk.

Binnen het onderwijsprogramma hebben studenten de mogelijkheid zich vroegtijdig te richten op de praktijk van de gezondheidszorg. Naast bepaalde stages die zij vanaf het eerste studiejaar volgen in de gezondheidszorg, zijn er ook vaardigheidstrainingen waaraan zij vrijwillig kunnen deelnemen. Deze trainingen worden verzorgd door het skillslab. De trainingen die gegeven worden, liggen op het gebied van de fysi-

sche diagnostiek, de therapie en laboratorium- en sociale vaardigheden. Een gedeelte van die vaardigheden is beschreven in de vorm van zogenaamde standaarden (Lodewick, 1978). Dit zijn met foto's en tekeningen aangevulde beschrijvingen van de verschillende stappen en handelingen, waaruit een vaardigheid bestaat.

2.5.7 Toetsing.

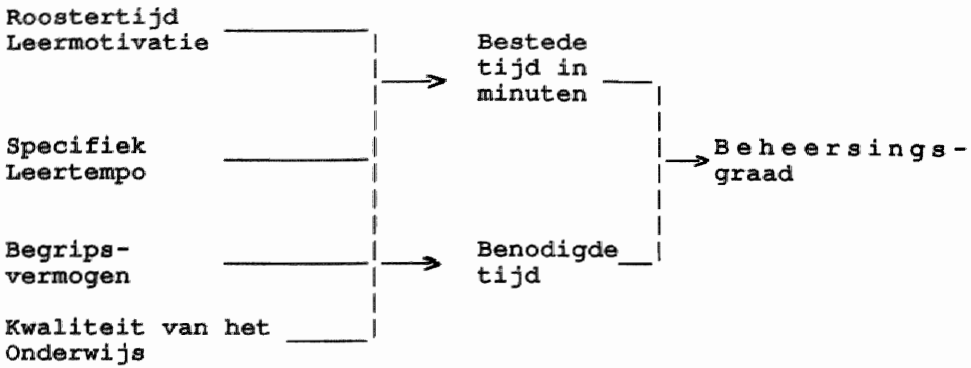
Na afloop van ieder blok wordt aan studenten een bloktoets voorgelegd. Deze toets bevat vragen over de inhoud van het blok. De bloktoets is in eerste instantie bedoeld om studenten feed-back te geven over hun leerresultaten in het afgelopen blok. Naast de bloktoets, die bedoeld is om kennis te meten die verworven is in het voorafgaande blok, bestaat er ook een zogenaamde voortgangstoets. In deze toets zijn vragen opgenomen die betrekking hebben op de eindtermen van de basisartsopleiding. Studenten worden dus geconfronteerd met vragen die de basisarts zou moeten kunnen beantwoorden. Deze toets wordt vier keer per jaar afgenomen. De faculteit heeft een aantal criteria ontwikkeld om op basis van deze toets beslissingen te nemen over de studievoortgang van studenten (Verwijnen et al. 1982).
Er is ook een toets ontwikkeld om de vaardigheden van studenten te meten. Deze toets, de vaardighedentoets, confronteert studenten met een aantal praktijksituaties waarin zij hun vaardigheden moeten demonstreren.

2.6 Enkele onderwijskundige theorieën over het onderwijs-leerproces.

In het voorgaande hebben wij ons vooral geconcentreerd op de beschrijving van relevante kenmerken van probleemgestuurd onderwijs, zoals die vorm hebben gekregen in het medische curriculum van de Rijksuniversiteit Limburg. Doel daarvan was aan te geven welke elementen noodzakelijkerwijze deel moeten uit maken van een meetinstrument om een dergelijk programma te evalueren. In deze paragraaf richten wij ons op meer algemene theorieën over leren en onderwijzen, eveneens met het doel om aan deze theorieën inzichten te ontlenen die bruikbaar kunnen zijn voor de evaluatie van dat programma.
Een algemene en empirisch getoetste theorie van het onderwijzen en leren, dat wil zeggen een theorie die geldig is voor alle leerkrachten, leerlingen, vakgebieden en onderwijssituaties, ontbreekt tot dusverre (De Corte, 1977; Haertel, Walberg & Weinstein, 1983).
Wel zijn er de afgelopen jaren pogingen ondernomen om modellen te ontwikkelen die de relatie tussen het onderwijsproces en het leerresultaat beschrijven (bijvoorbeeld Carroll, 1963; Bruner, 1964; Kounin, 1973; Bloom, 1976; Cooley & Leinhardt, 1976). Deze modellen richten zich vaak op zeer verschillende aspecten van onderwijs. Ze kunnen ingedeeld worden naar de eenheid van analyse waarop hun beschrijving zich richt (individu, klas, school), het domein waarop de theorie betrekking heeft (cognitief, psychomotorisch, affectief) of naar de mate van specificatie van de gebruikte concepten.
Modellen worden hier opgevat als hulpmiddelen voor onderzoek, ze beschrijven relaties tussen een aantal essentiële componenten in verband met de bestudeerde werkelijkheid. Het belang van modellen schuilt in het feit dat zij de werkelijkheid, zij het meestal sterk vereenvoudigd, beschrijven waardoor het bijvoorbeeld mogelijk wordt prescripties af te leiden voor het praktisch handelen.
Haertel, Walberg en Weinstein (1983) maken onderscheid tussen drie soorten modellen, namelijk: (1) modellen waarvan de belangrijkste constructen gedefinieerd zijn in termen van tijdsbesteding van leerlingen, en die gericht zijn op leren in schoolse situaties; (2) modellen gebaseerd op psychologische leertheorieën; (3) modellen die niet de noodzakelijke condities voor leren beschrijven op individueel niveau, maar zich richten op randvoorwaarden die noodzakelijk zijn voor het optimaal functioneren van onderwijs (bv. school- en klasorganisatie).
Het antwoord op de vraag welke modellen het beste gebruikt kunnen worden om bepaalde onderwijssituaties te beschrijven en te onderzoeken, wordt voornamelijk bepaald door de opvattingen die men heeft over het menselijk leren; en door het antwoord op de vraag welke doelstellingen nagestreefd worden met de beschrijving van dat onderzoek en de setting waarin het leren plaatsvindt. In dit proefschrift beperken we ons tot de beschrijving van enkele belangrijke modellen uit de

eerste groep volgens de indeling van Haertel et. al. (1983). Het gaat daarbij om de opvattingen van Carroll (1963), Bloom (1974) en Cooley en Lohnes (1976) die zich richten op het leren van individuen in schoolse situaties. Leerpsychologische modellen (Bruner, 1964; Gagne, 1974) worden hier buiten beschouwing gelaten omdat deze zich meer met het leren van individuen onder sterk gecontroleerde omstandigheden bezighouden (Haertel, Walberg & Weinstein, 1983).
In Carroll's model over schools leren is studietijd de centrale variabele. Verschillen in aanleg tussen leerlingen worden bijvoorbeeld uitgedrukt in verschillen in benodigde leertijd voor bepaalde taken. Carroll (1963) is van mening dat studenten, allen het gewenste prestatiecriterium kunnen bereiken indien zij de beschikking hebben over een voldoende tijd om een leertaak af te ronden en als zij deze tijd daadwerkelijk gebruiken. Carroll tracht de mate van studiesucces te voorspellen op basis van drie factoren: 'perseverance' (de hoeveelheid tijd die de student bereid is te investeren in een taak), 'aptitude' (de taakspecifieke vaardigheden die de student met zich meebrengt in de leersituatie) en de intelligentie of verbale begaafdheid van de student. De matstudiesucces is volgens Carroll afhankelijk van de verhouding tussen twee factoren namelijk de netto bestede en de netto benodigde studietijd. De netto bestede studietijd wordt bepaald door de motivatie van de student (perseverance), maar zij wordt veelal wel begrensd door de ruimte die het onderwijssysteem in tijd toelaat. Deze factor wordt door Carroll 'opportunity' genoemd: de mate waarin de student van de onderwijsorganisatie de gelegenheid krijgt studietijd te investeren in zijn studietaak. Behalve door de taakspecifieke aptitude van de student wordt de netto benodigde studietijd bepaald door de onderwijskwaliteit. Bovendien heeft de student een bepaalde mate van intelligentie nodig om een studietaak te begrijpen en om optimaal van het geboden onderwijs te kunnen profiteren. Carroll's model brengt de basisfactoren, behalve intelligentie, in kaart die een cursusontwerper kan manipuleren om de kwaliteit van een cursus te verhogen. In figuur 2.1 is het model van Carrol weergegeven.

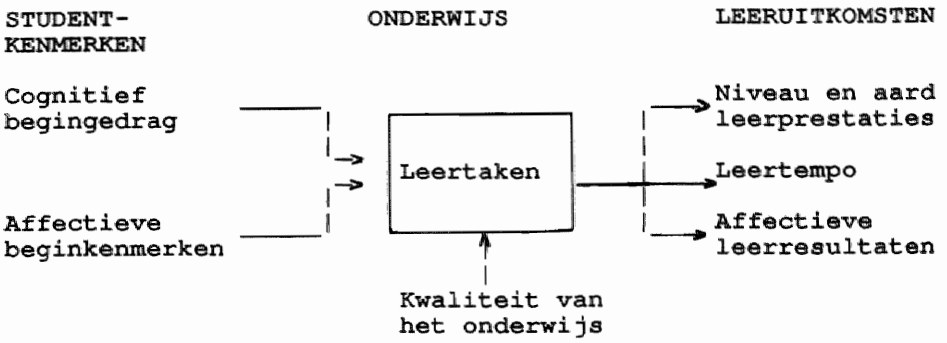
Figuur 2.1: Onderwijsleermodel volgens Carrol.



Carroll's werk gaf de aanzet tot de constructie van modellen die een aantal verfijningen bevatten (Bloom, 1974; Cooley & Lohnes, 1976; Harnischfeger & Wiley, 1976). Bloom (1974) bijvoorbeeld, constateerde in empirisch onderzoek grote verschillen tussen de bruto tijd die de student achter de boeken doorbrengt en de netto tijd die hij daadwerkelijk aan zijn taak besteedt (bijvoorbeeld ten gevolge van wegdromen, punten slijpen, even een praatje maken). Wiley en Harnischfeger (1974) richtten hun aandacht op systeembepaalde beperkingen van de feitelijk bestede studietijd. Zij constateerden, analoog aan Bloom, dat er onderscheid gemaakt moet worden tussen de bruto tijd die de student volgens het rooster zou kunnen besteden en de netto studietijd die er overblijft na aftrek van allerlei gedwongen verliestijden.
Om te weten hoeveel tijd leerlingen besteden aan zelfstudie is het niet voldoende na te gaan hoeveel tijd beschikbaar was in het programma. Tevens moet worden onderzocht hoeveel tijd feitelijk besteed is. Volgens Bloom (1974, 1976) is het rendement van het onderwijs niet alleen in belangrijke mate afhankelijk van het begingedrag van de leerlingen maar ook van de kwaliteit van de instructie. In zijn model is de kwaliteit van instructie afhankelijk van een aantal factoren. Er moeten aanwijzingen zijn in de leerstof over wat en hoe geleerd moet worden (cues), er moet van voldoende gelegenheid tot oefening (practice) geboden worden evenals positieve en negatieve bekrachtiging (reinforcement) en feedback, gevolgd door correctieve maatregelen.
Bloom onderscheidt in zijn model twee soorten begingedrag: het cognitieve en het affectieve. Het cognitieve begingedrag betreft de relevante voorkennis en de specifieke aanlegfactoren. De affectieve leerlingkenmerken hebben betrekking op de attitude van de leerling en interesse in de leertaak.
Er zijn drie soorten van leerresultaten in het model van Bloom: 1) het niveau en het type van de prestatie, 2) de

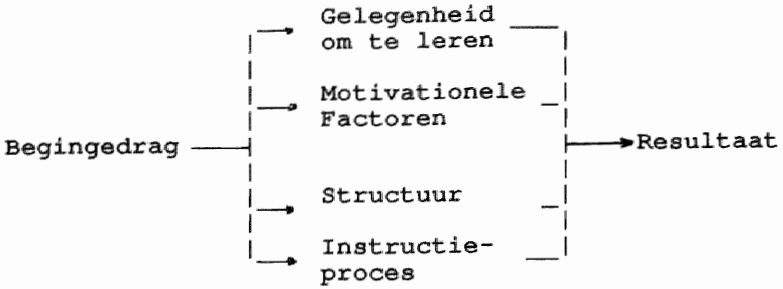
snelheid waarmee geleerd wordt, en 3) affectieve leerresultaten (verkregen interesse voor de leerstof). In figuur 2.2 is het model van Bloom weergegeven.

Figuur 2.2: Onderwijsleermodel volgens Bloom (1976).



Dit model is zowel descriptief als prescriptief. Men kan het enerzijds gebruiken om onderwijsleerprocessen mee te beschrijven, anderzijds kan het ook dienen als middel om richtlijnen te formuleren voor de vormgeving van het onderwijsleerproces. Bloom's overtuiging is dat maximalisering van de kwaliteit van instructie, automatisch leidt tot verhoging van de waarden van de drie door hem relevant geachte leeruitkomsten.
Cooley en Lohnes (1976) ontwikkelden een 'classroom-process'-theorie die gebaseerd is op de modellen van Glaser (1976), Gagne (1972) en Carroll (1963), en die zich evenals die welke we hiervoor bespraken, richt op de relaties tussen de onderwijspraktijk en het leerresultaat. Cooley en Lohnes veronderstellen, net als Carroll en Bloom, dat leren een resultaat is van het begingedrag van de leerlingen en van het onderwijsleerproces. Zij onderscheiden vier factoren in het onderwijsleerproces namelijk opportunity, motivators, structure, en instructional events. Opportunity is in navolging van Carroll gedefinieerd als de hoeveelheid tijd die studenten aan een leertaak kunnen besteden. Motivators zijn factoren die de interne en externe motivatie beïnvloeden. Onder interne motivatie wordt verstaan de motivatie die de leerling zelf heeft om met de taak bezig te blijven. Externe motivatie wordt beïnvloed door factoren als presentatie, reinforcement en feedback. De dimensie 'structure' duidt op de manier waarop het curriculum is opgezet (organisatie en sequentie van de leerstof). De dimensie 'instructional events' heeft te maken met de manier waarop de leraar met de leerlingen contacten legt (frequentie, duur, kwaliteit en inhoud). In figuur 2.3 is het model van Cooley en Lohnes weergegeven.

Figuur 2.3: Onderwijsleermodel volgens Cooley en Lohnes.



2.7 Beschrijving van de vragenlijst.

De in de vorige paragraaf beschreven modellen van Carrol (1963), Bloom (1976), en Cooley en Lohnes (1976) hebben als gemeenschappelijk kenmerk het zogenaamde 'input - throughput - output' karakter. Men definieert een beginsituatie en onderscheidt factoren die de start van het schoolse leerproces omschrijven; vervolgens worden relaties gelegd met factoren die betrekking hebben op het eigenlijke onderwijsleerproces, waarna relaties met 'onderwijsproducten' beschreven worden. Met behulp van deze modellen kan men proberen een aantal verschillende probleemstellingen te analyseren. Zo zou men een beschrijving van het onderwijsleerproces kunnen geven, of factoren op het spoor komen die men kan manipuleren om de kwaliteit van de output te verhogen. Ook bieden ze ons de gelegenheid om evaluatie-onderzoek te doen naar de werking van onderwijssystemen.
Een vergelijking van deze modellen laat zien dat zij een aantal centrale elementen bevatten:

- studentkenmerken (voorkennis, motivatie, studievaardigheden),
- docentkenmerken (doceervaardigheden),
- leerstofkenmerken (structuur van de leerstof, moeilijkheidsgraad),
- zelfstudiekenmerken (hoeveelheid bestede studietijd),
- programmakenmerken (verroosterde tijd, volgorde van leereenheden),
- productkenmerken (op affectief of cognitief gebied of m.b.t. vaardigheden).

Bij de constructie van de vragenlijsten (studenten- en tutorenversie) voor de beoordeling van het onderwijs in de geneeskunde, is zoveel mogelijk met deze elementen rekening gehouden. Er zijn vragen opgenomen over de aansluiting van de leerstof op de voorkennis van studenten, de moeilijkheidsgraad van de leerstof, de structuur van de leerstof, de opbouw van het blok, het gedrag van de docent, het aantal uren dat men besteed heeft aan zelfstudie, en de aansluiting

van de bloktoets op de bestudeerde leerstof. Kenmerken van bovengenoemde elementen zijn, voor zover mogelijk, vertaald naar de situatie van probleemgestuurd medisch onderwijs.
De vragenlijst voor studenten is opgezet naar een ontwerp van H.G. Schmidt (1981). Deze lijst is in de periode 1981/1984 verschillende malen bijgesteld. De eerste versie van deze lijst is in het academiejaar 1981/1982 onderzocht op zijn betrouwbaarheid en validiteit (Gijselaers, 1983). Naar aanleiding van dit onderzoek en op basis van gesprekken met docenten, studenten en leden van de projectgroep, is deze versie voor het academiejaar 1982/1983 op een aantal punten bijgesteld (zie bijlage 1). Wijzigingen in de daaropvolgende academiejaren betroffen slechts enkele itemformuleringen (zie eveneens bijlage 1). De vragenlijst voor tutoren heeft een vergelijkbare opzet als de vragenlijst voor studenten (zie bijlage 2).
De vragenlijst voor studenten is onderverdeeld in een aantal categorieën. Deze hebben betrekking op:
- de Algemene Indruk
- het Blokboek
- de Onderwijsgroep
- de Tutor
- de Leermiddelen
- het Skillslab
- de Inhoudsdeskundigen
- de Bloktoets
- de Afsluitende open vragen.

De categorieën algemene indruk, blokboek, onderwijsgroep, en tutor, bevatten ieder ongeveer 10 items. De andere categorieën omvatten ieder 3 - 4 items. In de meeste categorieën zijn items opgenomen die van beschrijvende aard zijn. In sommige categorieën zijn ook items opgenomen die een globaal beoordelend karakter hebben. In verschillende categorieën zijn zowel positief als negatief geformuleerde items opgenomen. De meeste items zijn van het zogenaamde 'Likert-type'. Deze items bevatten een bewering waarmee studenten het 'volledig oneens', 'tamelijk oneens', 'neutraal', 'tamelijk eens', of 'volledig eens' kunnen zijn. De antwoordcategorieën verwijzen naar de cijfers 1, 2, 3, 4, 5 die respectievelijk corresponderen met de bovengenoemde antwoordmogelijkheden 'volledig oneens' tot en met 'volledig eens'.

De laatste categorie omvat een tweetal open vragen (positieve- en negatieve aspecten van het blok), een vraag naar het aantal uren besteed aan literatuurstudie, en een 'overall rating item' (geef een globaal oordeel in de vorm van een schoolcijfer van 1 t/m 10 voor dit blok). In tabel 2.6 is de vragenlijst weergegeven zoals die in het academiejaar 1984/1985 werd gebruikt. Deze vragenlijst zal het uitgangspunt vormen voor verdere bespreking in het proefschrift.

Tabel 2.6: Vragenlijst voor de beoordeling van het onderwijs in de geneeskunde (academiejaar 1984/-1985).

Inleiding:

De fakulteit tracht verschillende gegevens te verzamelen voor de beoordeling van de doeltreffendheid van het onderwijs. Een zeer belangrijk gegeven is Uw oordeel over de gang van zaken in de verschillende blokken. Dat oordeel wordt gebruikt voor het bijstellen van het blok, zodat medestudenten uit latere jaren van uw ervaringen kunnen profiteren. In een enkel geval zal Uw oordeel voor de planningsgroep of de vakgroep Onderwijsontwikkeling en Onderwijsresearch aanleiding zijn nader kontakt met U op te nemen voor verduidelijking of specificering. De resultaten zullen gerapporteerd worden aan alle bij het onderwijs betrokkenen: planningsgroep, tutoren, onderwijscommissie, skillslab, etc..

Enkele opmerkingen over de vragenlijst zelf:
Het merendeel van de vragen bestaat uit beweringen, waarop U kunt reageren door omcirkeling van een cijfer.

Omcirkeling van het getal

1 betekent dat U het "volledig oneens" bent met de bewering;
2 betekent "tamelijk oneens";
3 betekent "neutraal", "er tussen in";
4 betekent "tamelijk eens";
5 betekent "volledig eens".

Als U de bewering echt niet van toepassing acht, dan slaat U de vraag over.
Let op, dat Uw mening gevraagd wordt over datgene wat zich volgens U in feite heeft afgespeeld, niet wat zich had moeten afspelen.
Aan het einde van de vragenlijst vindt U enkele open vragen; steekwoordsgewijze beantwoording is voldoende. Deze open vragen kunnen soms gevolgd worden door vragen die door de planningsgroep zijn geformuleerd.

We hopen op U bereidwillige medewerking.

Wim Gijselaers, vakgroep Onderwijsontwikkeling en Onderwijsresearch.

BLOKNUMMER: .....

ONDERWIJSGROEPSNUMMER:.....

(EVENTUEEL EXAMENNUMMER OF NAAM):.....)

	volledig oneens				volledig eens
ALGEMENE INDRUK.					
1. Over het geheel genomen heb ik de afgelopen periode prettig gewerkt.	1	2	3	4	5
2. Het blok sloot goed aan op mijn voorkennis.	1	2	3	4	5
3. De doelstellingen van dit blok waren mij duidelijk.	1	2	3	4	5
4. Het programma vergde veel studietijd.	1	2	3	4	5
5. De onderwerpen in dit blok waren volgens mij nuttig in het kader van de medische studie.	1	2	3	4	5
6. Het programma van dit blok heeft mijn attitude t.a.v. de gezondheidszorg beïnvloed.	1	2	3	4	5
7. De leerstof van dit blok was moeilijk.	1	2	3	4	5
8. Ik heb in dit blok veel opgestoken.	1	2	3	4	5
9. Ik vond de in dit blok aangeboden leerstof interessant.	1	2	3	4	5
HET BLOKBOEK.					
10. Informatie m.b.t. opzet en werkwijze was duidelijk gepresenteerd.	1	2	3	4	5
11. De taken waren voor het merendeel duidelijk omschreven.	1	2	3	4	5
12. De taken leenden zich voor het merendeel voor een systematische aanpak.	1	2	3	4	5

13. De taken waren zo sterk gestructureerd dat er weinig lol aan te beleven was.	1	2	3	4	5
14. De taken gaven voldoende aanleiding tot een zinnige groepsdiscussie.	1	2	3	4	5
15. De taken gaven voldoende aanleiding tot zelfstudie.	1	2	3	4	5
16. Ik heb in dit blok in belangrijke mate onafhankelijk van het blokboek gestudeerd.	1	2	3	4	5
17. De taken gaven voldoende aanknopingspunten voor het formuleren van leerdoelen.	1	2	3	4	5
18. Er was een grote varieteit aan onderwerpen in het blokboek.	1	2	3	4	5
19. Er was een grote variëteit aan taken in het blokboek.	1	2	3	4	5
20. Er was een grote variëteit aan hulpmiddelen bij de taken in het blokboek (videobanden, dia's, foto's, etc.).	1	2	3	4	5
21. De taken waren aanleiding om onderwerpen uit basisvakken te bestuderen (anatomie, biochemie, etc.).	1	2	3	4	5
22. Door het werken met de taken in het blokboek was het mogelijk de blokdoelstellingen te realiseren.	1	2	3	4	5
23. De hoeveelheid taken in het blokboek moet worden uitgebreid.	1	2	3	4	5

ONDERWIJSGROEP.

24. De onderwijsgroep maakte gebruik van systematische werkprocedures bij het aanpakken van de taken.	1	2	3	4	5
25. Het werken in de groep betekende een stimulans voor mijn zelfstudie-activiteiten.	1	2	3	4	5

26. In de onderwijsgroep werden steeds duidelijk afspraken gemaakt m.b.t. de te bestuderen stof.	1	2	3	4	5
27. Iedereen hield zich aan zijn afspraken.	1	2	3	4	5
28. Ik heb de bijeenkomsten als prettig ervaren.	1	2	3	4	5
29. De bijeenkomsten waren productief.	1	2	3	4	5
30. Iedereen leverde een aktieve bijdrage.	1	2	3	4	5
31. Ik heb de onderwijsgroepsbijeenkomsten als een rem ervaren in de voortgang van mijn studie.	1	2	3	4	5
32. Ik heb in dit blok in belangrijke mate onafhankelijk van de leerdoelen van de onderwijsgroep gestudeerd.	1	2	3	4	5

TUTOR.

33. De tutor gaf blijk een goed begrip te bezitten van de doelstellingen van het blok.	1	2	3	4	5
34. De tutor leek op de hoogte van de onderwijskundige uitgangspunten van het onderwijssysteem.	1	2	3	4	5
35. De tutor gaf de indruk zijn/haar rol plezierig te vinden.	1	2	3	4	5
36. De tutor stimuleerde tot hard werken.	1	2	3	4	5
37. De tutor stelde regelmatig discussie-stimulerende vragen.	1	2	3	4	5
38. De tutor stuurde regelmatig met zijn eigen vakkennis de discussie.	1	2	3	4	5
39. De tutor stimuleerde het maken van afspraken m.b.t. de te bestuderen leerstof.	1	2	3	4	5
40. De tutor controleerde het nakomen van gemaakte afspraken.	1	2	3	4	5

41. De tutor stimuleerde het raadplegen van inhoudsdeskundigen. 1 2 3 4 5

42. De tutor stimuleerde het gebruik maken van andere leer- en evaluatiemiddelen. 1 2 3 4 5

43. Regelmatig evalueerde de tutor met ons de gang van zaken in de onderwijsgroep. 1 2 3 4 5

44. De tutor functioneerde over het geheel genomen goed in zijn/haar rol als tutor. 1 2 3 4 5

SKILLSLAB-VAARDIGHEDEN.

45. Ik vond de training(en) lichamelijk onderzoek in dit blok zinvol (incl. patiëntenkontakten). 1 2 3 4 5

46. Ik vond de training(en) therapeutische vaardigheden in dit blok zinvol. 1 2 3 4 5

47. Ik vond de training(en) laboratoriumvaardigheden in dit blok zinvol. 1 2 3 4 5

48. Ik ben tevreden over de begeleiding bij bovengenoemde trainingen. 1 2 3 4 5

49. Ik vond de training(en) sociale vaardigheden in dit blok zinvol. 1 2 3 4 5

50. Ik vond de simulatiepatiënten-kontakten in dit blok zinvol. 1 2 3 4 5

51. Ik ben tevreden over de begeleiding bij de training(en) sociale vaardigheden c.q. simulatie-patiënten-nabespreking. 1 2 3 4 5

LEERMIDDELEN.

52. In het studielandschap was een voldoende verscheidenheid aan literatuur beschikbaar. 1 2 3 4 5

53. Er waren voldoende av-middelen beschikbaar. 1 2 3 4 5

54. De beschikbare av-middelen waren inhoudelijk van goede kwaliteit.	1	2	3	4	5

INHOUDSDESKUNDIGEN.

55. Onze onderwijsgroep heeft een aantal malen een inhoudsdeskundige geraadpleegd, namelijk: AANKRUISEN HETGEEN VAN TOEPASSING IS	0	1-2	3-4	5-6	meer dan 6 keer

FORMATIEVE EVALUATIE.

56. De bloktoets sloot aan bij de door mij bestudeerde onderwerpen.	1	2	3	4	5
57. De bloktoets toetste onderwerpen die in de doelstellingen omschreven waren.	1	2	3	4	5
58. De zelfevaluatiemiddelen sloten aan bij de inhoud van het blok.	1	2	3	4	5

ENKELE AFSLUITENDE OPEN VRAGEN.

59. Hoeveel tijd hebt u, gemiddeld genomen, per week aan literatuurstudie besteed?

60. Als u het totale onderwijs zoals u het in dit blok meemaakte, globaal genomen, een cijfer zou moeten geven op schaal 1 tot 10 (6 is een voldoende), welk cijfer geeft u dan?

61. Welke aspecten of onderdelen van dit blok zouden veranderd moeten worden?

62. Welke aspecten of onderdelen van dit blok vond U erg goed?

(c) Vakgroep Onderwijsontwikkeling en Onderwijsreseach RL

HOOFDSTUK 3. BETROUWBAARHEID EN VALIDITEIT VAN STUDENTOORDELEN: THEORETISCHE OVERWEGINGEN.

3.1 Inleiding.

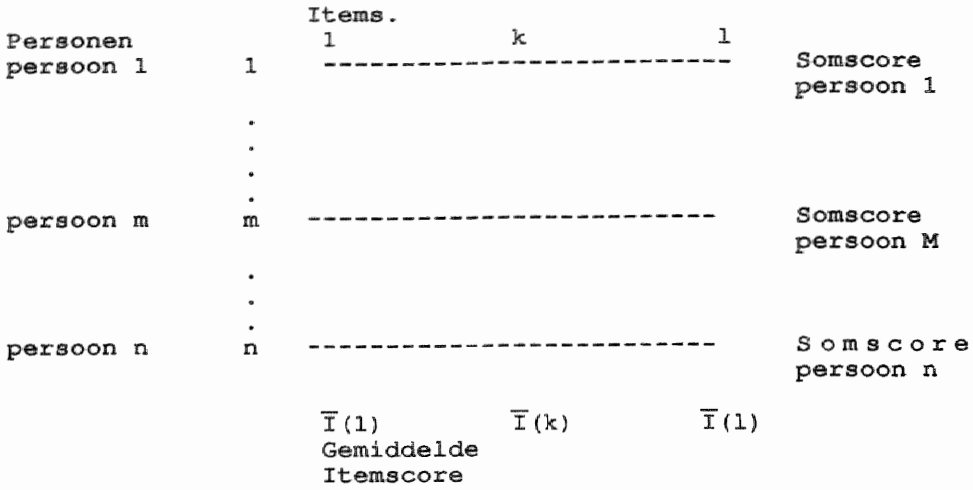
In hoofdstuk 2 is aandacht besteed aan de constructie van beoordelingsvragenlijsten. In het bijzonder werd aandacht besteed aan de constructie van de beoordelingsvragenlijst die binnen de medische faculteit gebruikt wordt. Een voor de hand liggende vraag is in hoeverre dergelijke instrumenten daadwerkelijk geschikt zijn voor evaluatiedoeleinden. Dit betreft met andere woorden de vraag naar de betrouwbaarheid en validiteit van beoordelingsvragenlijsten. In onderzoek in deze werden veelvuldig tegenstrijdige resultaten gevonden (Cohen, 1981; Dowell & Neal, 1982; Abrami, Leventhal & Perry, 1982; Marsh, 1984). Rodin en Rodin (1972) toonden bijvoorbeeld aan dat studentoordelen geen validiteit bezitten. Anderen daarentegen hebben herhaaldelijk resultaten gevonden die juist wel op de betrouwbaarheid en validiteit van studentoordelen wijzen (Cohen, 1981; Marsh, 1982; Marsh & Hocevar, 1984). Over het geheel genomen kan volgens de meeste onderzoekers echter geconcludeerd worden dat studentoordelen als een betrouwbare en valide indicator voor de kwaliteit van het onderwijs beschouwd kunnen worden, mits de gehanteerde beoordelingsvragenlijsten aan een aantal voorwaarden voldoen (multidimensionele structuur van de lijst, merendeels bestaande uit beschrijvende items, etc.).
In dit hoofdstuk gaat het om de vraag hoe de betrouwbaarheid en validiteit van beoordelingsvragenlijsten onderzocht kan worden. In hoofdstuk 4 zal namelijk empirisch onderzoek beschreven worden, dat binnen de medische faculteit verricht is naar de betrouwbaarheid en validiteit van de in hoofdstuk 2 beschreven vragenlijst. In dit hoofdstuk zal eerst een conceptueel model beschreven worden voor metingen die gebaseerd zijn op oordelen. Aan de hand van dit model worden vervolgens enkele problemen toegelicht die zich met betrekking tot de betrouwbaarheid en validiteit van oordelen kunnen voordoen. Daarna worden enkele, in de praktijk veelvuldig gebruikte, methoden gepresenteerd waarmee de betrouwbaarheid en validiteit van studentoordelen onderzocht kan worden. Tenslotte wordt een literatuuroverzicht gegeven met betrekking tot onderzoek naar de betrouwbaarheid en validiteit van studentoordelen.

3.2 Betrouwbaarheid en validiteit van beoordelingsvragenlijsten: een conceptueel model voor metingen met behulp van oordelen.

Het gebruik van beoordelingsvragenlijsten berust op de aanname dat mensen in staat zijn om op een objectieve en nauwkeurige manier kenmerken van een of ander object waar te nemen, en er een of andere "waarde" (value) aan te hechten, die in relatie staat tot de kenmerken van dat object (De afbeelding van de werkelijkheid in een getalsysteem). Mensen worden in zo'n geval als instrument gebruikt om metingen te verrichten. Een dergelijk meetinstrument brengt een aantal problemen met zich mee. Men moet bijvoorbeeld hopen dat mensen zich in hun oordeel niet laten beïnvloeden door allerlei vooroordelen over de werkelijkheid. Als dat zo zou zijn, kan men verwachten dat de nauwkeurigheid en objectiviteit van de meting afneemt. Dit zijn vragen die betrekking hebben op de betrouwbaarheid en validiteit van oordelen. De eerste vraag betreft de meetprecisie van het instrument. De tweede vraag heeft betrekking op de correspondentie tussen het meetresultaat en datgene wat het instrument bedoelt te meten.
Carmines en Zeller (1979) omschrijven betrouwbaarheid als de mate waarin de meetuitkomsten consistent zijn: hoe hoger de consistentie bij herhaalde metingen, hoe hoger de betrouwbaarheid. Een instrument is betrouwbaar als het bij herhaling, onder gelijke omstandigheden, dezelfde meetresultaten produceert. Validiteit wordt gedefinieerd als de mate waarin een instrument werkelijk datgene meet wat men beoogt te meten. Cronbach (1972) omschrijft validiteit als de mate waarin een meetinstrument geschikt is voor de doeleinden waarvoor het gebruikt wordt of gaat worden. Hij koppelt dus in tegenstelling tot Carmines en Zeller (1979), het begrip validiteit expliciet aan de doeleinden waarvoor meetresultaten gebruikt worden. De omschrijvingen van Carmines en Zeller (1979) en Cronbach (1972) worden in de literatuur veelvuldig gebruikt om de begrippen betrouwbaarheid en validiteit te onderscheiden. Het voert in het kader van dit proefschrift te ver om nader op de discussies omtrent de definiëring van beide begrippen in te gaan. Als uitgangspunt voor de definiering van het begrip betrouwbaarheid volgen we de opvatting van Carmines en Zeller (1979). Validiteit wordt omschreven conform de definitie van Cronbach (1972). Deze definitie is in de context van het in dit proefschrift beschreven onderzoek beter bruikbaar dan de definitie van Carmines en Zeller (1979). Hier gaat het immers met name om de vraag in hoeverre studentoordelen gebruikt kunnen worden om de kwaliteit van onderwijs te verbeteren.
Methoden om de betrouwbaarheid en validiteit van oordelen te bepalen, berusten in sterke mate op de uitgangspunten van de klassieke testtheorie. De klassieke testtheorie is in eerste instantie ontwikkeld om de betrouwbaarheid en validiteit van eendimensionele psychologische tests te bepalen. Dit zijn

instrumenten die slechts een enkelvoudig kenmerk of een enkelvoudige trek van een individu meten (bijvoorbeeld extraversie of introversie). De klassieke uitgangspunten van de testtheorie zijn desondanks bruikbaar voor de bepaling van de betrouwbaarheid en validiteit van oordelen. De analysemethoden die binnen deze theorie ontwikkeld zijn, kunnen echter niet klakkeloos op beoordelingsvragenlijsten toegepast worden. Deze vragenlijsten verschillen namelijk op een aantal punten van tests. Tests worden gebruikt om kenmerken van een individu te meten. Dit gebeurt door personen een verzameling items voor te leggen die alle afkomstig zijn uit een homogeen itemdomein: een verzameling items die een enkele trek, een enkel kenmerk of een enkele dimensie meten.
Ieder item meet het te onderzoeken kenmerk partieel. De items kunnen op deze wijze als partieel herhaalde metingen beschouwd worden. Kenmerken van individuen worden meestal beoordeeld door deze individuen zelf: zij doen 'uitspraken' op items over een aspect van hun eigen persoonlijkheid. De itemscores van de test worden gesommeerd tot een totale testscore. De somscore is een maat voor de mate waarin een bepaald kenmerk bij een bepaalde persoon voorkomt. De betrouwbaarheid van een test die dat kenmerk meet, is in hoge mate afhankelijk van de mate waarin items als herhaalde metingen van elkaar beschouwd kunnen worden.
Beoordelingsvragenlijsten worden hier gedefinieerd als instrumenten die kenmerken meten buiten de persoon: oordelen over een object, een gebeurtenis of een andere persoon etc. Ze hebben vaak een multidimensionaal karakter, omdat verschillende dimensies van een object als het ware "tegelijkertijd" gemeten worden. Een voor de hand liggend voorbeeld is de in dit proefschrift beschreven vragenlijst. Deze vragenlijst wordt verondersteld verschillende aspecten van probleemgestuurd onderwijs te meten. De items van een multidimensionale beoordelingsvragenlijst zijn dus niet afkomstig uit een enkel homogeen itemdomein en hoeven dat ook niet te zijn. Dat betekent dat niet alle items van een dergelijke vragenlijst als replicaties van elkaar gezien kunnen worden. Oordelen worden <u>per item</u> gesommeerd en gemiddeld (en dus niet <u>over</u> items, zoals bij psychologische tests). Op basis van de berekende scores per item worden uitspraken gedaan over het beoordeelde object.
Het verschil tussen tests en beoordelingsvragenlijsten kan gedemonstreerd worden aan de hand van de onderstaande datamatrix.

Figuur 3.1: Datamatrix personen x items.



Een test is zodanig geconstrueerd dat de items of subgroepen van items bepaalde facetten van een construct meten. De somscore van een persoon over de items van een test duidt de mate aan waarin die persoon het kenmerk onder beschouwing bezit. Eventuele verschillen in scores met andere personen worden toegewezen aan verschillen in de mate waarin die personen dat kenmerk bezitten. De betrouwbaarheid en validiteit van de test wordt bepaald door eigenschappen van items. Bij beoordelingsvragenlijsten echter gaat het om de vraag in hoeverre een item, of groepen items, op een betrouwbaar en valide wijze kenmerken van een object anders dan de persoon die de vragenlijst invult, meten. In dit geval bepalen de kwaliteiten of eigenschappen van de personen voornamelijk de betrouwbaarheid en validiteit van de vragenlijst. Enkel als personen voldoende homogeen oordelen, vindt men een hoge betrouwbaarheid. Men zou kunnen zeggen dat de testtheorie in bovenstaande figuur uitspraken probeert te doen over de rijen van de matrix en dat een meettheorie voor oordelen uitspraken doet over de kolommen van de matrix. Dat wil overigens niet zeggen, dat de klassieke testtheorie onbruikbaar zou zijn voor de beoordelingssituaties waarover wij hier spreken. Integendeel verderop zal blijken dat deze theorie zeer wel toepasbaar is, zij het dan op een "gekantelde" matrix. Het wil ook niet zeggen dat we in dit proefschrift verder alleen nog maar geïnteresseerd zijn in de homogeniteit van oordelen over enkelvoudige items. Zoals zal blijken kunnen groepen items (bijvoorbeeld items met betrekking tot het blokboek of onderwijsgroep) gebruikt worden om schalen te creëren die een weerspiegeling vormen van de factoren die bepalend zijn voor de kwaliteit van probleemgestuurd medisch onderwijs.

Om de betrouwbaarheid en validiteit van beoordelingsvragenlijsten te bepalen, heeft men dus een meettheorie nodig die de psychometrische kenmerken van beoordelaars eerder dan die van items in ogenschouw neemt. Een dergelijke theorie wordt beschreven door Guilford (1954). Deze theorie maakt gebruik van operationele definities over betrouwbaarheid en validiteit, die ontleend zijn aan de klassieke testtheorie. De meettheorie van Guilford (1954) is op het terrein van studentoordelen met name verder ontwikkeld door Feldman (1977), Larson (1979) en Marsh (1984). We zullen in het kort de definities uit de klassieke testtheorie omschrijven, alvorens de (klassieke) theorie van Guilford (1954) en de theorie van Larson (1979) te beschrijven.
De klassieke testtheorie veronderstelt dat een gemeten testscore (geobserveerde score) bestaat uit de som van de ware score en een errorscore of meetfout (Lord & Novick, 1968). Deze relatie kan als volgt in een eenvoudig additief model weergegeven worden:

$$O = W + E$$

O = geobserveerde score
W = ware score
E = errorscore, toevallige meetfout

De ware score wordt in dit model gezien als een gemiddelde score die tot stand komt, onder identieke omstandigheden en bij elkaar niet beïnvloedende metingen, wanneer die meting een groot aantal malen herhaald zou kunnen worden (De Groot & Van Naerssen, 1977). De term identiek impliceert dat het mogelijk is om herhaaldelijk metingen bij een individu te verrichten zonder dat het individu verandert, een conditie die in de werkelijkheid uiteraard niet realiseerbaar is. Bij het bepalen van de betrouwbaarheid veronderstelt men dat de errorscore een toevallige meetfout is die tot stand komt door onnauwkeurigheden in de meting (bijvoorbeeld t.g.v. slordigheden van diegene die de meting uitvoert, slecht geconstrueerd meetinstrument, of slordig invulgedrag van proefpersonen). De ware score is niet gerelateerd aan de errorscore.
Toevalsfluctuaties in de geobserveerde score tasten de betrouwbaarheid aan van het meetinstrument en zijn te wijten aan een bepaalde proportie error als onderdeel van de geobserveerde score. Naarmate de geobserveerde score meer afhankelijk is van de errorscore, neemt de betrouwbaarheid van een meting af. De klassieke testtheorie veronderstelt dat bij een groot aantal metingen de gemiddelde geobserveerde score gelijk zal zijn aan de gemiddelde ware score (de toevallige meetfouten worden verondersteld elkaar op te heffen met als gevolg dat de gemiddelde errorscore nul wordt).

In de klassieke testtheorie geldt dat de variantie van de geobserveerde scores gelijk is aan de som van de varianties van de ware scores en de errorscores: Var O = Var W + Var E.

Met behulp van deze vergelijking kan de betrouwbaarheid van een meetinstrument als volgt gedefinieerd worden: de betrouwbaarheid van een test is gelijk aan de ratio van de variantie van de ware scores en de variantie van de geobserverveerde scores. Deze ratio wordt de betrouwbaarheidscoëfficient genoemd.

$$\text{Betrouwbaarheids-coëfficiënt} = \frac{\text{Variantie ware scores}}{\text{Variantie geobserveerde scores}} = \frac{\text{Var w}}{\text{Var o}}$$

Inspectie van deze definitie toont dat naarmate de foutenvariantie kleiner is, de variantie van de geobserveerde scores, de variantie van de ware scores benadert (Lord & Novick, 1968). De betrouwbaarheidscoëfficiënt nadert dan de waarde 1.0. Naarmate er sprake is van meer errorvariantie zal de coëfficient een lagere waarde aannemen. De meting is dan minder betrouwbaar.
De meettheorie van Guilford (1954) over oordelen maakt, analoog aan de klassieke testtheorie, onderscheid tussen geobserveerde score, ware score en errorscore. Met de notatiewijze van Larson (1979) zijn de relaties tussen deze scores als volgt in een additief model weer te geven. Het model heeft de volgende vorm:

$$X_O = X_W + E$$

X_O = een geobserveerd oordeel over een object
X_W = het ware oordeel over een object
E = de totale errorscore in het geobserveerd oordeel

De errorscore wordt in het model gesplitst in twee componenten: nonsystematische error en systematische error. Beide errorcomponenten zijn niet gerelateerd aan de ware score. De nonsystematische error fluctueert volgens een random patroon en beïnvloedt derhalve de betrouwbaarheid van het oordeel. De systematische error volgt een bepaald patroon over beoordelaars, objecten of een combinatie van beide en beïnvloedt derhalve de validiteit van het oordeel. De relatie tussen geobserveerde score, ware score, random error en systematische error krijgt dan de volgende vorm:

$$X_O = X_W + S + R$$

X_O = geobserveerde score
X_W = ware score
S = systematische error
R = random error

De systematische errorcomponent kan weer onderverdeeld worden in 'contingente systematische error' en 'noncontingente systematische error'. De noncontingente systematische error bestaat uit errorsoorten als 'response set error of leniency' (neiging van individuen om de schaaluitersten van items te gebruiken bij hun oordeel), 'response set error of central tendency' (neiging van individuen om vooral het schaalmidden van items te kiezen) en 'restriction of range' (neiging van individuen om vaak dezelfde waarden te gebruiken). Noncontingente error treedt vooral dan op wanneer een beperkte set van individuen systematisch een bepaald invulgedrag vertonen, onafhankelijk van datgene waarop het meetinstrument betrekking heeft. Deze vorm van error heeft een constant effect op de meting. Het effect is niet afhankelijk van kenmerken van het beoordeelde object. Contingente error is daarentegen afhankelijk van kenmerken van het beoordeelde object. De twee belangrijkste vormen van contingente error zijn volgens Guilford (1954): het 'halo effect' en de 'logical error'. Hij beschrijft het halo-effect als volgt: '.... We judge our fellows in terms of a general mental attitude toward them; and there is, dominating this mental attitude toward the personality as a whole, a like mental attitude toward particular qualities.' Logical error wordt door Guilford (1954) gedefinieerd als: '.. error due to the fact that judges are likely to give similar ratings for traits that seem logically related in the minds of the raters. ' In de literatuur over meten met behulp van oordelen zijn uiteenlopende definities van beide errorsoorten te vinden. Het voert in het kader van deze studie te ver om op deze discussies nader in te gaan. De lezer wordt in dat verband verwezen naar twee uitstekende overzichtsartikelen: respectievelijk van Cooper (1981) en van Saal, Downey en Lahey (1980).

In de literatuur over studentoordelen worden beide errorsoorten samengevat onder de term 'implicit theories of behavioral covariation' (Larson, 1979). Hiermee doelt men op het verschijnsel dat beoordelaars soms impliciete aannamen doen over relaties tussen bepaalde aspecten van een object, die van invloed kunnen zijn op het oordeel over dat object. Als mensen het idee hebben dat gedrag x gepaard gaat met gedrag y, dan zal men geneigd zijn een oordeel te geven dat behelst dat gedrag y daadwerkelijk is opgetreden als gedrag x waargenomen werd.

In ons onderzoek kunnen dergelijke fenomenen bijvoorbeeld voorkomen als studenten impliciete aannamen hebben over het functioneren van onderwijsgroepen. Stel dat zij het idee hebben dat het slecht functioneren van onderwijsgroepen te wijten is aan bepaald tutorgedrag (b.v. geen interesse voor het blok), dan zal men geneigd zijn een oordeel te geven dat behelst dat men dat gedrag waarnam bij het constateren van een slecht functionerende onderwijsgroep.

Het onderscheid tussen contingente- en noncontingente systematische error krijgt in het model de volgende vorm:

$$X_o = X_w + IT_{x|y} + NC + R$$

X_o = geobserveerde score
X_w = ware score
$IT_{x|y}$ = impliciete theorie over de covariantie tussen gedrag x en y
NC = noncontingente systematische error
R = random error

Larson (1979) maakt bij de impliciete theorieën die in het beoordelingsproces een rol kunnen spelen, nog een onderverdeling naar normatieve theorieën en idiosyncratische theorieën. Normatieve theorieën worden door iedereen in een bepaalde populatie gedeeld. Idiosyncratische theorieën zijn gebonden aan individuele beoordelaars.
Uiteindelijk kan men metingen op basis van oordelen als volgt in een model gieten:

$$X_o = X_w + AN_{x|y} + AI_{x|y} + NC + R$$

X_o = Geobserveerde score
X_w = Ware score
$AN_{x|y}$ = Normatieve impliciete theorie over de covariantie tussen gedrag x en y
$AI_{x|y}$ = Idiosyncratische impliciete theorie over de covariantie tussen gedrag x en y
NC = Noncontingente systematische error
R = Random error

Het model van Larson (1979) is bruikbaar om onderzoek te verrichten naar de betrouwbaarheid en validiteit van studentoordelen. Het levert handvaten om op een zorgvuldige manier de kwaliteit van studentoordelen te onderzoeken.
In het bovenstaande model veronderstelt men dat oordelen functioneren als testitems. De betrouwbaarheid van de gemiddelde oordelen van een groep personen (het equivalent van de totale gemiddelde testscore), is een functie van de mate van overeenstemming tussen personen binnen een verzameling personen (het equivalent van de mate waarin de items hetzelfde kenmerk meten) en het aantal personen dat oordelen gaf (het equivalent van het aantal items van een test). Verschillen in oordelen tussen personen, binnen een groep, worden als random error beschouwd.
Betrouwbaarheid wordt in dit model eveneens gedefinieerd als de ratio tussen de variantie van de ware scores en de geobserveerde scores. De betrouwbaarheid wordt verlaagd naarmate er meer random error vertegenwoordigd is in de geobserveerde score. In de volgende paragrafen zal aandacht besteed worden

aan methoden om de betrouwbaarheid en validiteit van studentenoordelen te onderzoeken.

3.3 Methoden om de betrouwbaarheid van tests en beoordelingsvragenlijsten te schatten.

In onderzoek over de betrouwbaarheid van studentoordelen vindt men de concepten uit de klassieke testtheorie over de betrouwbaarheid van testen veelvuldig terug (Feldman, 1977; Marsh & Overall, 1979). Zij worden echter op een andere manier gebruikt. We zullen in het kort enkele vormen van testbetrouwbaarheid bespreken, waarna aandacht besteed wordt aan vormen van betrouwbaarheid van beoordelingsvragenlijsten.

3.3.1 Methoden om de betrouwbaarheid van tests te schatten.

De betrouwbaarheid van een test wordt berekend om inzicht te krijgen op de invloed van toevalsfluctuaties op de testscore. De meeste methoden berusten ofwel op A) het principe van herhaald testonderzoek, ofwel op B) de bewerking van een enkel testonderzoek. Goldstein en Hersen (1984) onderscheiden binnen het herhaald testonderzoek de volgende soorten betrouwbaarheid: A1) de paralleltest betrouwbaarheid en A2) de test-hertest betrouwbaarheid. Op basis van bewerkingen van een enkel testonderzoek maken zij onderscheid tussen: B1) de split-half betrouwbaarheid en B2) de interne consistentie van een test.

A1: Paralleltest-betrouwbaarheid.

Bij de betrouwbaarheid wordt de proportie ware variantie van een test geschat door een tweede test (die geacht wordt hetzelfde te meten) af te nemen, en de scores van beide testen te correleren. Als de tweede test dezelfde ware variantie heeft als de test waarvan men de betrouwbaarheid wil bepalen, en de scores van de tweede test alleen van de oorspronkelijke test afwijken door toevalsvariantie die in beide tests voorkomt, dan is de correlatie tussen de beide tests een schatting van de betrouwbaarheid van de oorspronkelijke test. Men noemt de op deze manier bepaalde correlatiecoëfficient ook wel de equivalentiecoëfficient.

A2: Test-hertest betrouwbaarheid.

Een andere vorm van herhaald testonderzoek is de test-hertest. Deze wordt berekend door dezelfde test na een bepaald tijdsinterval weer af te nemen. De correlatie tussen de individuele scores op beide testen is een maat voor de betrouwbaarheid van de test.

B1: Split-half betrouwbaarheid.

De split-half betrouwbaarheid van een test wordt bepaald door een test in twee delen van gelijke lengte te splitsen (eerste helft-tweede helft, of even items - oneven items). Men berekent de scores op beide delen en correleert deze met elkaar. De correlatiecoëfficient is een schatting voor de betrouwbaarheid van de halve test. Een schatting van de betrouwbaarheid van de hele test, kan men verkrijgen met behulp van de Spearman-Brown formule (Drenth, 1975). Met behulp van deze formule kan men de invloed van testverlenging op de betrouwbaarheid berekenen.
De split-half methode geeft een schatting van de interne consistentie van de test. Zij is echter gebaseerd op een enkele splitsing van de test, alle andere splitsingen die mogelijk zijn worden buiten beschouwing gelaten.

B2: Interne consistentie.

Er zijn twee alternatieve methoden ontwikkeld die de interne consistentie meten, om aan dit bezwaar tegemoet te komen. De test wordt niet in twee gelijke delen opgesplitst, maar in de afzonderlijke items. Op deze manier kan men de homogeniteit of interne consistentie van een test berekenen. Onder homogeniteit verstaat men de mate waarin de afzonderlijke items hetzelfde kenmerk of dezelfde eigenschappen meten. Procedures om de interne consistentie van een test te bepalen, veronderstellen dat ieder item een gedeelte van een enkel kenmerk meet en dat ieder item een bepaalde hoeveelheid random errorvariantie bezit. Aangezien random error geminimaliseerd kan worden door te middelen over items, is de betrouwbaarheid van de totale test afhankelijk van het aantal items en de mate waarin de items eenzelfde kenmerk meten. De homogeniteits-betrouwbaarheid kan berekend worden met behulp van de Kuder-Richardson 20 (KR-20) formule en de coëfficient alpha. De KR-20 formule is bedoeld voor tests waarin de items twee antwoordmogelijkheden hebben (bijvoorbeeld goed/fout of waar/onwaar). Coëfficient alpha wordt meestal berekend voor tests die testitems bevatten met meerdere antwoordmogelijkheden (bijvoorbeeld vijfpuntsschalen).

3.3.2 Methoden om de betrouwbaarheid van oordelen te schatten.

Er bestaan een aantal methoden om de betrouwbaarheid van oordelen, of beoordelingsvragenlijsten te bepalen. Deze berusten alle op het principe van het vaststellen van de mate van overeenstemming tussen personen over een object. Verschillen in oordelen tussen personen worden opgevat als random error die de betrouwbaarheid van oordelen verlaagt.
Een manier om de betrouwbaarheid van oordelen te bepalen is het berekenen van de interbeoordelaarscorrelatie. Individuele oordelen worden per item van de beoordelingsvragenlijst

met elkaar gecorreleerd. Hoe hoger de correlatie tussen de oordelen, hoe hoger de betrouwbaarheid van de oordelen.
Een andere veelgebruikte methode om de betrouwbaarheid van individuele oordelen te bepalen is het berekenen van de intraclass correlatiecoëfficient. Deze coëfficient geeft, in tegenstelling tot interbeoordelaars-correlatie, een beeld van de mate van overeenstemming tussen mogelijke paren beoordelaars binnen een categorie of groep. De intraclass correlatie kan ook beschouwd worden als een index voor de mate van relatieve homogeniteit tussen beoordelaars binnen een groep, ten opzichte van een verzameling groepen (Guilford, 1954; Tinsley & Weiss, 1975; Shrout & Fleiss, 1979). De intraclass correlatie is alleen bruikbaar in situaties waarin de betrouwbaarheid bepaald wordt van individuele oordelen die als metingen op interval-niveau beschouwd kunnen worden.
De intraclass correlatie wordt gewoonlijk berekend door de totale variantie op te splitsen in variantiecomponenten. Afhankelijk van het te kiezen variantie-analytisch design, de eenheid van analyse, en de vraag of men over items of personen wil generaliseren, worden verschillende formules voor de berekening van de intraclass correlatie toegepast (Shrout en Fleiss, 1979).
Een derde methode is het berekenen van de mate van absolute overeenstemming tussen personen. Men onderzoekt in hoeverre personen binnen een groep hetzelfde oordeel geven op een item. De absolute overeenstemming tussen personen geldt dan als een maat voor de betrouwbaarheid van oordelen.
Een andere benadering is de generaliseerbaarheidstheorie van Cronbach, Gleser, Nanda en Rajaratnam (1972). Met behulp van deze theorie kan men onderzoeken in hoeverre men over items of beoordelaars kan generaliseren naar andere groepen items of personenpopulaties.
Naarmate individuele oordelen meer gegeneraliseerd kunnen worden naar andere populaties, zijn oordelen meer betrouwbaar.

3.4 Methoden om de validiteit van tests en beoordelingslijsten te schatten.

In het model van Larson (1979) wordt de validiteit van oordelen aangetast door het optreden van systematische error. Naarmate deze errorcomponent een groter onderdeel vormt van de geobserveerde score, wordt de validiteit van de meting lager. De correspondentie tussen het meetresultaat en de kenmerken van het gemeten construct verdwijnt. Er bestaan verschillende manieren om de validiteit van een instrument te bepalen. We zullen hieronder drie vormen van validiteit bespreken die van belang zijn voor beoordelingsvragenlijsten: inhoudsvaliditeit, criteriumvaliditeit en constructvaliditeit.

1. Inhoudsvaliditeit.

De inhoudsvaliditeit van een instrument wordt bepaald door een oordeel te geven over de mate waarin de inhoud van de items waaruit dat instrument is samengesteld, representatief zijn voor datgene waarover conclusies getrokken moeten worden. De inhoudsvaliditeit van een instrument kan niet berekend worden, maar alleen vastgesteld worden aan de hand van oordelen van experts of proefpersonen over de inhoud van het instrument en de relatie met datgene wat het beoogt te meten. De constructie van een inhoudsvalide instrument vereist een zorgvuldige definiëring van datgene wat men wil meten.
Inhoudsvaliditeit is een belangrijke vorm van validiteit voor bijvoorbeeld onderwijskundige toetsen. Als kennis en vaardigheden die nodig zijn om een toets te kunnen maken, overeen komen met de kennis en vaardigheden zoals omschreven in het onderwijsprogramma, kunnen de scores van individuen op dergelijke toetsen (bijvoorbeeld reken- of leesvaardigheidstoetsen) gebruikt worden als indicatie van de mate waarin zij welomschreven kennis en vaardigheden verworven hebben (Meerling, 1981).
Inhoudsvaliditeit is echter ook een belangrijke vorm van validiteit voor beoordelingsvragenlijsten. Wil men inhoudsvalide beoordelingsvragenlijsten construeren, dan is het belangrijk dat er een analyse van het onderwijsprogramma gemaakt wordt, en algemeen onderwijskundig belangrijke variabelen in de lijst zijn opgenomen. De mate van overeenstemming tussen onderzoekers, over de inhoudelijke representativiteit van datgene wat een instrument beoogt te meten, bepaalt de inhoudsvaliditeit van het instrument.

2. Criteriumvaliditeit.

Deze vorm van validiteit is vooral belangrijk voor de validering van psychologische tests. Bij criteriumvaliditeit gaat het steeds om de relatie tussen een testscore of een reeks van testscores met een of meer criteriumvariabelen (Meerling, 1981). Een instrument is meer valide naarmate het sterker correleert met een extern criterium.
Bij de bepaling van de criteriumvaliditeit van beoordelingsvragenlijsten die de kwaliteit van onderwijs meten, gebruikt men vaak als criterium de studieprestaties van studenten. Predictive validiteit wordt bepaald door op basis van testscores in de toekomst gelegen criteriumgedrag te voorspellen.

Naarmate de correlatie met het criterium hoger is, is de predictieve validiteit van een test hoger.
Aan deze vorm van validatie zijn een aantal problemen verbonden. Het eerste probleem heeft betrekking op de keuze van het criterium.
Doordat de validatie sterk afhankelijk is van de correlatie tussen meetresultaten en criterium, rijst de vraag welk criterium gekozen moet worden, een belangrijke vraag die geen

eenvoudig antwoord kent. Vandaar dat Carmines en Zeller (1979) opmerken dat er geen 'single criterion-related validity coëfficiënt' bestaat. Er bestaan zoveel validiteitscoëfficiënten, als er criteria beschikbaar zijn. Een bijkomend probleem is dat, naarmate een instrument abstractere concepten meet, het moeilijker is om een geschikt criterium te vinden.

3. Constructvaliditeit.

Deze vorm van validiteit is ontwikkeld uit onvrede met het valideren van instrumenten met behulp van inhoudsvaliditeit of criterium validiteit (Cronbach & Meehl, 1955). Beide vormen van validiteit creëren een aantal problemen die hiervoor al voldoende aangeduid zijn: inhoudsvaliditeit is afhankelijk van de mate van consensus tussen experts, en criterium-validiteit staat of valt met de kwaliteit en relevantie van het gekozen criterium. De APA (American Psychological Association) Committee on Psychological Tests ontwikkelde in de periode 1950-1954 een validatiemethode die men construct validiteit noemde. Uitgangspunt was dat als een test ontworpen is om een of ander latent construct te meten, men in staat moet zijn om te onderzoeken welk latent construct gemeten wordt. Een latent construct (bijvoorbeeld intelligentie) is een concept dat gebruikt wordt om kenmerken van individueel gedrag te kunnen beschrijven. Men kan dit gedrag niet op een directe wijze bepalen of meten, anders zou men geen test nodig hebben. Maar men heeft wel een theorie ter beschikking over het latente construct, een theorie die beschrijft hoe het latente construct onder bepaalde omstandigheden functioneert. In de theorie vindt men aanknopingspunten voor de constructie van de test, en de theorie geeft aanwijzingen hoe men de kwaliteit van de test kan bepalen.
Het proces van constructvalidering komt sterk overeen met het proces van verificatie of falsificatie van een theorie. De bepaling van de construct-validiteit van een instrument verloopt in drie stappen. De eerste stap heeft betrekking op de specificatie van theoretische relaties tussen de variabelen of concepten. De tweede stap berust op het empirisch onderzoek naar de relaties tussen de variabelen of concepten zoals die in de eerste stap gehypothetiseerd zijn. De laatste stap betreft het zoeken naar een antwoord op de vraag in hoeverre de onderzoeksresultaten de hypotheses bevestigen of ermee in strijd zijn.
Er bestaan verschillende technieken om de constructvaliditeit van een instrument te bepalen: de multitrait/multimethodbenadering, de confirmerende factoranalyse en de methode van structurele vergelijkingen. Campbell en Fiske (1959) beschrijven een techniek die gebruik maakt van een zogenaamd multitrait/multimethod design. In een dergelijk design wordt een 'trait' of persoonlijkheidstrek met behulp van verschillende instrumenten gemeten (inclusief het te valideren instrument). Tevens worden 'traits' gemeten die geen theoretische relatie mogen hebben met de oorspronkelijke 'trait'. De

correlatiematrix die ontstaat op basis van correlaties tussen de verschillende metingen over de verschillende 'traits', noemt men een multitrait/multimethod correlatiematrix. De validiteit van een instrument is dan afhankelijk van hoge correlaties met instrumenten die geacht worden hetzelfde te meten (convergerende validiteit), en lage tot nul correlaties met instrumenten die traits meten die geen relatie hebben met het te valideren instrument (discriminerende validiteit). Een andere techniek is gebaseerd op het postuleren van een bepaalde factor structuur van het te valideren instrument. Met behulp van confirmerende factoranalyse kan men bepalen in hoeverre een gepostuleerde structuur afwijkt van de empirisch gevonden factorstructuur (Goldstein & Hersen, 1984). Een derde techniek is gebaseerd op de methode van structurele vergelijkingen. Men postuleert bepaalde relaties tussen variabelen en onderzoekt met behulp van structurele vergelijkingen in hoeverre de gepostuleerde structuur afwijkt van de empirische structuur.

3.5 De betrouwbaarheid van studentoordelen: een literatuuroverzicht.

De betrouwbaarheid van studentoordelen is over het algemeen op drie manieren onderzocht: 1) op basis van interbeoordelaars betrouwbaarheid, 2) op basis van interne consistentie van oordelen, en 3) op basis van de stabiliteit van de oordelen (zie o.a.: Costin, Greenough & Menges, 1971; Kulik & McKeachie, 1975; Feldman, 1977; Blom & Langerak, 1979; Centra, 1979; Marsh & Overall, 1979; Marsh, 1984).
In deze paragraaf zullen deze methoden en de resultaten ervan kort besproken worden.

1. Interbeoordelaarsbetrouwbaarheid.

Feldman (1977) beschrijft een aantal veelgebruikte procedures om de interbeoordelaarsbetrouwbaarheid van studentoordelen te bepalen. Binnen iedere procedure wordt de mate van overeenstemming tussen beoordelaars op een of een verzameling items onderzocht. De mate van overeenstemming is gelijk aan de betrouwbaarheid van oordelen.
Een methode is gericht op het correleren van individuele oordelen op items van beoordelingsvragenlijsten. Men trekt at random studentenparen binnen een cursusgroep en correleert de oordelen van deze paren op de afzonderlijke items. Deze procedure werd onder andere gevolgd door Bausell en Magoon (1972a, 1972b, 1972c). De betrouwbaarheid van individuele oordelen varieerde rondom .30. De betrouwbaarheid van groepsoordelen (ongeveer 30 studenten) daarentegen lag rondom .90 per beoordelingsitem.
Een vergelijkbare methode is het splitsen van klassen in twee gelijke groepen. Deze methode is gebaseerd op de split-half betrouwbaarheid van tests: men splitst (in het kader van de "averechtse" benadering) niet een verzameling items in twee

helften maar een groep studenten. Men berekent de gemiddelde groepscores op een item of verzameling items (die een schaal vormen), en correleert deze met elkaar. Met behulp van de Spearman - Brown formule voor testverlenging krijgt men een schatting van de betrouwbaarheid van de oordelen van alle studenten in een klas, maar ook een schatting van de betrouwbaarheid van oordelen op individueel niveau (nl. als men de formule ten behoeve van testverkorting toepast).
Feldman (1977) beschrijft studies van Bausell en Magoon (1972c; 1972d), Gillmore (1973), Guthrie (1949; 1954), Hoyt (1969; 1973a; 1973b), Maslow en Zimmerman (1956), Murray (1975), en Remmers en Weisbrodt (1964) waarin de betrouwbaarheden op deze wijze berekend werden. De betrouwbaarheden van oordelen op klasniveau (ongeveer 30 studenten) lagen tussen .70 en .90.

Twee andere methodes benaderen het vraagstuk van betrouwbaarheid op variantie-analytische wijze. In de eerste benadering vergelijkt men de variantie binnen groepen met de variantie tussen groepen en berekent de intra-class correlatie (Winer, 1962). Op deze wijze verkrijgt men een schatting van de relatieve homogeniteit van de oordelen. Een lage intraclass correlatie wijst op een grote mate van variatie binnen groepen. Een hoge intraclass correlatie wordt verkregen als de variantie tussen groepen met verschillende docenten groter is dan de variantie binnen groepen met een enkele docent.
Centra (1973) pastte de intraclass-procedure toe op een vragenlijst die 30 items bevatte. Hij varieerde tevens het aantal studenten dat de lijst invulde. Bij een aantal van tien studenten vond hij voor de meeste items een intraclass correlatie van .70 en .78 voor oordelen die betrekking hadden op het totaal functioneren van docenten. Bij een aantal van respectievelijk vijftien en twintig studenten steeg de intraclass correlatie van .80 naar .90.
Marsh (1982) berekende eveneens de intraclass correlatie van studentoordelen met wisselende groepsgrootte. Afhankelijk van de groepsgrootte bedroeg de intraclass correlatie, .95 bij 50 studenten, .90 bij 25 studenten, .74 bij een groepsgrootte van 10 studenten, en .23 voor oordelen op individueel niveau. Marsh (1982) concludeerde op basis van deze resultaten dat studentoordelen betrouwbaar zijn, mits gebaseerd op minimaal tien studenten.
Een andere benadering, die echter weinig wordt toegepast, wordt beschreven door Gillmore, Kane en Naccarato (1978). Zij bepalen de betrouwbaarheid van studentoordelen met behulp van de generaliseerbaarheidstheorie van Cronbach, Gleser, Nanda, en Rajaratnam (1972). Deze theorie houdt rekening met een veelheid van factoren (de beoordelaar, de testitems, tijdstip van beoordeling) die van invloed kunnen zijn op de errorvariantie. Gillmore, Kane , en Naccarato (1978) onderzochten hoeveel studentoordelen nodig zijn om een betrouwbaar oordeel over het optreden van een docent te krijgen in het geval dat men studentoordelen laat meewegen bij personele kwesties als

bevorderingen of ontslag van docenten. Zij laten zien dat studentoordelen in dergelijke gevallen betrouwbaar zijn als die betrekking hebben op vijf cursussen van een docent, waarbij de klassegrootte een omvang moet hebben van 15 studenten of meer.

2. Interne consistentie van oordelen.

Soms gebruikt men een procedure die gebaseerd is op de interne consistentie van individuele oordelen. Men onderzoekt in hoeverre studenten consistent zijn in hun oordeel over een docent. Een vragenlijst wordt eenmalig afgenomen, waarbij de studenten een docent op een aantal aspekten beoordelen. Bij dit type onderzoek gaat het meestal om de vraag in hoeverre doceervaardigheid, opgevat als een eendimensionaal kenmerk, betrouwbaar te meten is. Deze procedure behandelt de items van een beoordelingsvragenlijst als items van een test. Items worden als replicaties van elkaar gezien, die een eendimensionale trek meten. Deze benadering kan gebruikt worden om de homogeniteit van afzonderlijke dimensies binnen een vragenlijst te bepalen.
Marsh (1982b) berekende coëfficient alpha, om de homogeniteit te bepalen van groepen items die deel uitmaken van een aantal factoren van de door hem ontwikkelde beoordelingsvragenlijst (SEEQ). Hij rapporteert voor de verschillende factoren alpha's die varieren tussen .88 en .97.
Deze methode is echter onbruikbaar voor het bepalen van de totale betrouwbaarheid van een multidimensionale vragenlijst. Een ander probleem is dat deze methode impliceert dat oordelen van studenten toch betrouwbaar kunnen zijn ondanks tegengestelde opvattingen over de kwaliteit van een cursus. Verschillen tussen studenten worden, in tegenstelling tot de interbeoordelaars-betrouwbaarheid, niet gezien als een bron van onbetrouwbaarheid. De coëfficient alpha zal daarom een overschatting geven van de betrouwbaarheid van studentoordelen.

3. Stabiliteit van studentoordelen.

Bij een derde methode om de betrouwbaarheid van studentoordelen te bepalen, wordt de stabiliteit van de studentoordelen berekend (docenten worden met een zeker tijdsinterval twee of meerdere malen beoordeeld).
De stabiliteit (equivalent aan de test-hertest betrouwbaarheid) kan bepaald worden door studenten na een bepaald tijdsinterval, opnieuw het onderwijs te laten beoordelen met behulp van dezelfde vragenlijst. Remmers (1959) verrichte zo'n onderzoek naar de stabiliteit van studentoordelen. Hij verzamelde 7974 beoordelingsvragenlijsten (de zogenaamde Purdue Rating Scale) met betrekking tot 112 docenten. Drie weken later werd dezelfde lijst nog eens door dezelfde groep studenten ingevuld. De correlatie tussen beide metingen bedroeg .95. Kohlan (1973) vond in een vergelijkbaar onder-

zoek echter een correlatie van .58 tussen studentoordelen gemeten na het tweede cursusuur, en oordelen gemeten in het laatste cursusuur. Centra (1979) vond een gemiddelde correlatie van .70, waarbij 296 docenten tweemaal werden beoordeeld, de eerste keer direct na de cursus en de tweede keer vijf weken later. In een longitudinaal onderzoek van Marsh en Overall (1979) beoordeelde eenzelfde groep studenten (totaal 850 studenten) met hetzelfde beoordelingsinstrument het onderwijs na afloop van de cursus en een jaar later. De correlatie tussen de gemiddelde oordelen na afloop van de cursus (berekend op groepsniveau, N = 85 groepen) en de oordelen na een kalenderjaar bedroeg .81. Marsh en Overall (1979) berekenden met behulp van de Spearman-Brown formule tevens de stabiliteitscoëfficiënt op individueel niveau. Deze bedroeg .59. Zij concludeerden dat deze resultaten aantoonden dat individuele studentoordelen opvallend betrouwbaar (d.i. stabiel) waren.
Marsh (1977) onderzocht de stabiliteit van studentoordelen met behulp van cross-sectionele methoden. Aan studenten uit een jaargroep die afstudeerde, werd verzocht aan te geven welke docenten (met betrekking tot leersituaties in klasverband) een wezenlijke of juist een minimale bijdrage hadden geleverd aan hun leerervaringen. Docenten werden binnen een cursus als 'goed' beoordeeld als zij minimaal drie oordelen 'goed' kregen en tevens drie maal zoveel goede beoordelingen kregen als slechte beoordelingen. Op deze wijze verkreeg men een overzicht van 'goede' en 'slechte' docenten. Het daaropvolgende academiejaar werden dezelfde docenten opnieuw, op de gebruikelijke wijze, door een andere jaargroep studenten beoordeeld. Uit de resultaten bleek dat docenten die eerst als goed of slecht beoordeeld waren, ook een vergelijkbare beoordeling kregen door de volgende jaargroep studenten. Marsh (1977) concludeerde dat studenten onafhankelijk van elkaar, met hetzelfde instrument, gelijkluidende oordelen over docenten hadden gegeven. Dit is opnieuw een aanwijzing voor de stabiliteit van studentoordelen.

3.6 De betrouwbaarheid van studentoordelen:interpretatie en conclusies.

In de vorige paragrafen zijn verschillende methoden beschreven om de betrouwbaarheid van studentoordelen te bepalen. Deze richtten zich op: 1) vormen van interbeoordelaarsbetrouwbaarheid, 2) homogeniteit van items, en 3) de stabiliteit van oordelen.
Uit onderzoek naar de interbeoordelaars-betrouwbaarheid blijkt dat studentoordelen betrouwbaarder zijn naarmate meer studenten een oordeel geven (Feldman, 1977; Centra, 1979; Marsh, 1982a, 1984), een verschijnsel analoog aan het verschijnsel dat een test bestaande uit veel items over het algemeen betrouwbaarder is dan een test bestaande uit weinig items. Feldman (1977) wijst er daarom op dat één studentoordeel een kleine betrouwbaarheid heeft (.20). Studentoordelen

zijn volgens hem pas voldoende betrouwbaar als men zich baseert op een minimum aantal van 20 tot 25 studenten. Marsh (1982a) hanteert als criterium een minimum aantal van tien studenten. Er kleven een aantal problemen aan het bepalen van de betrouwbaarheid van studentoordelen op basis van de interbeoordelaars- betrouwbaarheid. Betrouwbaarheid wordt in de meeste studies opgevat als consistentie in oordelen tussen studenten. De aanname is dat de oordelen onafhankelijk van medestudenten tot stand komen, en dat de oordelen als replicaties beschouwd kunnen worden. Feldman (1977) betwijfelt echter de correctheid van deze assumptie. Volgens hem hoeft variantie tussen beoordelaars binnen een groep, niet altijd te wijzen op random error. De mogelijkheid bestaat immers dat bepaalde studentkenmerken zoals geslacht, aantal jaren schoolervaring en voorkennis met betrekking tot een cursus van invloed kunnen zijn op de wijze waarop het onderwijs beoordeeld wordt. Hij vond echter geen ondersteuning voor deze hypothese. Een ander probleem met het onderzoek naar de interbeoordelaars-betrouwbaarheid van studentoordelen vormt de vraag of men de feitelijke, of absolute overeenstemming tussen beoordelaars moet berekenen, of de relatieve overeenstemming tussen beoordelaars. De hierboven beschreven onderzoeken richtten zich niet op absolute overeenstemming, d.w.z. op de vraag of studenten binnen een groep hetzelfde oordeel geven op een item, maar richtten zich op de mate waarin de oordelen covarieren c.q correleren, de relatieve overeenstemming dus.
Een tweede conclusie is dat de items van subschalen van beoordelingsvragenlijsten meestal vrij homogeen zijn (Marsh, 1982a, 1982b). Marsh (1982, 1984) wijst er echter op dat dergelijke maten geen rekening houden met gebrek aan overeenstemming tussen beoordelaars. Vandaar dat de homogeniteit van beoordelingsvragenlijsten een overschatting van de betrouwbaarheid geeft. Tenslotte werd in een aantal onderzoeken gevonden dat de oordelen over een bepaalde tijdsperiode stabiel zijn (Remmers, 1959; Kohlan, 1973; Page, 1974; Marsh, 1977, 1979; Marsh & Overall, 1979, Centra, 1979).

3.7 De validiteit van studentoordelen: een literatuuroverzicht.

De meest frequente kritiek op studentoordelen is dat deze niet voldoende valide zouden zijn. Daarmee bedoelt men dat studenten in hun oordeel beïnvloed worden door factoren die feitelijk niet gerelateerd zijn aan de kwaliteit van het genoten onderwijs (zijnde het construct dat men wil meten), maar door factoren die daarvoor irrelevant zijn.
Eerder in dit hoofdstuk werd het model van Larson (1979) beschreven. In dit model werd onderscheid gemaakt tussen systematische error en nonsystematische error. Beide vormen van error zijn in dit model van invloed op de validiteit van oordelen. In de literatuur over studentoordelen spreekt men in het geval van systematische error ook over "bias"-bronnen (bias= "onzuiverheid" of "systematische fout"): factoren die de validiteit van oordelen op een negatieve wijze beïnvloeden. Naar de invloed van bias-bronnen is vrij veel onderzoek verricht in laboratorium- en veldstudies. Onderzoek naar de criterium- en constructvaliditeit van studentoordelen bestaat voornamelijk uit veldonderzoek.
Het bias-onderzoek naar de validiteit van studentoordelen is gericht op het ontdekken van externe factoren (of bronnen van bias) die een invaliderende invloed hebben op het oordeel van studenten over docenten. Factoren die volgens sommigen een rol zouden kunnen spelen, zijn: de achtergrondkenmerken van studenten (geslacht, sociaal-economische afkomst), het behaalde of verwachte tentamenresultaat, de grootte van de cursusgroep, de zwaarte van de kursus, de aard van de cursus (verplicht of facultatief), en de vriendelijkheid van de docent.
Een ander type onderzoek is gericht op de relatie tussen studentenoordelen en indicatoren voor goed onderwijs. Dit onderzoek tracht de criteriumvaliditeit van studentoordelen te bepalen door studentoordelen te correleren met studieprestaties, zelfbeoordelingen van docenten, oordelen van visitatiecommissies over docenten, de aard van de wetenschappelijke publikaties van docenten, en oordelen van collegadocenten.
Een laatste vorm van onderzoek is gebaseerd op de construct-validering benadering. Binnen dit type onderzoek tracht men met verschillende instrumenten en verschillende beoordelaars verschillende dimensies te meten van het construct 'kwaliteit van onderwijs'. In z'n algemeenheid kan men zeggen dat deze manier om de validiteit van studentoordelen te bepalen, bestaat uit onderzoek naar de relatie tussen de geobserveerde score en andere indicatoren van het te meten construct.
We zullen in de volgende paragrafen laboratorium- en veldstudies naar de validiteit van studentoordelen beschrijven.

3.7.1 Laboratorium-studies naar de validiteit van studentoordelen.

In dit gedeelte worden enkele laboratoriumstudies beschreven. Laboratoriumonderzoek heeft zich voornamelijk gericht op de vraag of studenten in hun oordeel over docenten, meer beinvloed worden door de persoonlijke stijl van de docent of door de didactische vaardigheden. Aan dit type onderzoek ligt het idee ten grondslag dat studentoordelen beinvloed kunnen worden door het charisma, of persoonlijkheid van de docent. In een experimenteel onderzoek van Naftulin, Ware en Williams (1973) werd namelijk geconstateerd dat een geestige, enthousiaste docent blijkbaar de oordelen van toehoorders op een positieve manier beinvloedde. Dat wil zeggen, de oordelen waren aanmerkelijk positiever dan men op grond van de inhoud van de lezing zou mogen verwachten. Dit resultaat leidde tot twijfels over de waarde van oordelen over docenten.
Wat was het geval? In het kader van een onderzoek naar de validiteit van studentoordelen, liet men een nepdocent (Dr. Fox genaamd), een geschoold acteur, een lezing geven voor een gehoor van psychiaters, psychologen en andere werkers in de geestelijke gezondheidszorg. Dr. Fox hield een lezing over een onderwerp waarmee de meeste toehoorders niet vertrouwd waren. Hij was geïnstrueerd een vlot en grappig verhaal te houden over dit onderwerp, waarbij het geheel op een geestige en enthousiaste manier gepresenteerd werd. De inhoud van de lezing berustte voornamelijk op louter nonsens. Na afloop van de lezing werd de toehoorders verzocht een vragenlijst in te vullen over de kwaliteit van de lezing. De items hadden zowel betrekking op de inhoud als vorm van de lezing. Ondanks het feit dat de lezing inhoudelijk weinig voorstelde, gaven de toehoorders een positief oordeel over inhoud en vorm van de lezing. Een enkeling merkte zelfs op dat hij behoefte had aan meer informatie over het onderwerp. De conclusie van de onderzoekers was dan ook, dat zelfs ervaren luisteraars zich blijkbaar in hun oordeel over de doceerkwaliteit laten beinvloeden door het geestig optreden van een docent. Hetgeen met andere woorden betekende, dat studenten dan nauwelijks in staat geacht konden worden om een valide oordeel te geven. Uit de reacties op dit onderzoek bleek echter dat de conclusies van de onderzoekers nauwelijks gehandhaafd konden worden (Marsh & Ware, 1982). Een gebrekkige methodologische opzet (kortere lezing dan gebruikelijk in het onderwijs, geen controlegroep, geen studenten in de onderzoeksgroep), lieten volgens deze critici dergelijke conclusies niet toe.
Als reactie op deze kritiek verrichtten Ware en Williams (1977) een aantal studies waarin gemanipuleerd werd met de variabelen 'doceervaardigheden' en 'inhoud van de lezing'. In deze studies werden verschillende lezingen op videoband opgenomen. De lengte van de lezingen werd constant gehouden, door extra voorbeelden en herhalingen toe te voegen, waarbij enthousiaste en saaie docenten hetzelfde verhaal hielden. De onderzoekers kwamen weer tot de conclusie dat studentoordelen

wel beïnvloed werden door de expressiviteit van de docent, maar niet door de inhoud.
Latere onderzoeken waarin deze resultaten opnieuw geanalyseerd werden lieten echter zien dat de beoordeling van de doceerstijl niet gerelateerd was aan de inhoud van de lezing (Marsh & Ware, 1982). Hetgeen erop wijst dat studenten blijkbaar wel onderscheid kunnen maken tussen de kwaliteit van de inhoud van de lezing en de kwaliteit van de doceervaardigheid van docenten.
Vergelijkbare studies zijn door Nisbett en Wilson (1976), en door Van Rooyen en Vlaander (1983) verricht. In deze studies werd de invloed van halo-effecten op studentoordelen onderzocht. Van Rooyen en Vlaander (1983) verrichtten twee experimenten waarin de invloed van halo-effecten op studentoordelen over docenten werd onderzocht. In het eerste experiment werden aan studenten videobanden getoond waarin een docent een 'warme rol' speelt (aandacht en interesse voor studenten) en een 'koude rol' (afstandelijk, wantrouwig). Vervolgens werd aan studenten gevraagd zich in evaluatieve oordelen (aardig-onaardig, interessant-oninteressant) uit te drukken over de verwachtingen die zij hadden over de colleges van de docent. Uit dat onderzoek kwam naar voren dat studenten verwachten dat de docent die een warme rol speelde interessantere colleges zal geven dan de docent die de koude rol speelde. In het tweede experiment werden dezelfde videobanden aan studenten getoond. Studenten kregen de opdracht het uiterlijk voorkomen, de gesticulatie en het accent van de docent te beoordelen aan de hand van beoordelingsvragen die geen evaluatief maar een descriptief karakter hadden. De resultaten van dit experiment gaven aan, dat persoonskenmerken van de docent (warme of koude rol) geen invloed lijken te hebben op de beoordeling van het uiterlijk voorkomen, accent en gesticulatie. De onderzoekers concludeerden derhalve dat studenten wel beinvloed worden in hun sympathie voor docenten, maar niet in de beschrijving van docentgedrag.

3.7.2 Veldstudies: Onderzoek naar de criterium-validiteit van studentoordelen.

De criteriumvaliditeit van een instrument (in ons geval: studentoordelen over de kwaliteit van het onderwijs) wordt vastgesteld door deze te relateren aan een of meer externe variabelen die beschouwd kunnen worden als "andere" goede operationalisaties van het te meten construct (in dit geval: de kwaliteit van het gegeven onderwijs). Het probleem dat zich dan natuurlijk voordoet, is dat onduidelijk is welk criterium of welke criteria een goede operationalisatie vormt of vormen van datgene wat men met behulp van studentoordelen beoogt te meten. Gebrek aan overeenstemming tussen experts over wat kenmerken van goed onderwijs zijn, heeft ertoe geleid dat soms geheel verschillende criteria gebruikt zijn (Cohen, 1981). Als studentoordelen bijvoorbeeld betrekking hebben op docentengedrag, is het belangrijk te weten wat

kenmerken zijn van goede docenten. McKeachie (1979) definieerde effectief docentengedrag als: ' the degree to which an instructor facilitates student achievement'. Als men deze definitie accepteert is het nog maar een kleine stap om af te leiden dat de mate waarin een student in een kursus iets heeft geleerd - zijn studieprestatie-, een criterium kan zijn voor effectief docentengedrag. Studieprestaties of toetsresultaten worden daarom regelmatig als criteriumvariabelen voor de effecten van goed onderwijs voorgesteld (Bijvoorbeeld: Bendig, 1953; McKeachie, Lin, & Mann, 1971; Rodin & Rodin, 1972; Frey, Leonard, & Beatty, 1975; Centra, 1977; Braskamp, Caulley, & Costin, 1979; Marsh & Overall, 1980). Andere criterium-variabelen die soms gebruikt worden zijn zelfevaluaties van docenten, beoordelingen door andere docenten ("class-room visitation" "peer evaluation"), en zelfs de onderzoeks-produktiviteit van docenten, uitgedrukt in aantallen publikaties. Van bovenstaande criteriumvariabelen zal in dit verband alleen aandacht besteed worden aan de relatie tussen studentoordelen en studieprestaties. Andere criteriumvariabelen blijken namelijk niet noemenswaardig samen te hangen met studentoordelen (Kulik & McKeachie, 1975; Centra, 1979; Marsh, 1984).
Studentoordelen worden in de V.S. met name gebruikt voor beoordelingen van docenten. Andere aspecten van het onderwijs, zoals de kwaliteit van de leermiddelen of de zwaarte van de cursus komen nauwelijks aan bod. De vragenlijsten bevatten voor het merendeel items die betrekking hebben op doceergedrag van docenten. Vandaar dat in het hieronder besproken onderzoek voornamelijk gesproken wordt over studentoordelen over docenten, veeleer dan over studentoordelen met betrekking tot de kwaliteit van het onderwijs.
Validiteitsstudies met als criterium "studieresultaten" of "toetsprestaties" worden meestal als volgt opgezet:

- 1. Er worden beoordelingsvragenlijsten gebruikt die behalve beschrijvende items ook evaluerende items bevatten (zoals: 'geef een globaal oordeel over de docent' of 'geef een cijfer voor deze kursus als geheel').
- 2. Als criterium voor studieresultaten worden meestal toetsresultaten gebruikt, zoals behaald op de kursus die door studenten beoordeeld werd.
- 3. Binnen een enkele kursus kunnen meerdere docenten worden beoordeeld, bijvoorbeeld als een zeer grote groep studenten de cursus volgt. Een dergelijke groep wordt dan opgesplitst in een aantal kleinere die les krijgen van assistenten van de verantwoordelijke docent. Als de totale groep studenten opgesplitst is in een aantal klassen, waarbij iedere klas steeds les krijgt van dezelfde assistent, spreekt men van een "multisection" kursus. Onderzoeken gebaseerd op multisection kursussen heten "multisection validiteitsonderzoeken".

- 4. Aanwijzingen voor de validiteit van studentoordelen kunnen verkregen worden door aan te tonen dat de klassen die de docent beter beoordeelden dan de andere klassen, ook betere toetsresultaten behaalden. Voorwaarde is dat het cursusmateriaal, het examen en de cursusopzet in alle secties hetzelfde is.

De aanname is dat verschillen in prestaties tussen groepen veroorzaakt kunnen worden door verschillen in de kwaliteit van het doceergedrag van docenten, wanneer sprake is van een voor iedere groep gelijke kursusopzet. Marsh en Overall (1980) beschrijven een dergelijk design als volgt:
'..each section of the course should be taught by a separate instructor, but the course outline, textbooks, course objectives and final examination should be developed by a course director who does not actually lecture to the students'. Een hoge correlatie tussen studentoordelen en toetsresultaten beduidt in deze opvatting dus dat studentoordelen validiteit bezitten. Er bestaan verschillende overzichtsartikelen waarin de resultaten van studies, die betrekking hebben op multisection kursussen met elkaar vergeleken worden. (bijvoorbeeld Page, 1974; Kulik & Kulik, 1974; Feldman, 1976; Centra, 1979; Marsh, 1980, 1984; Cohen, 1981; Dowell & Neal, 1982). We zullen een paar van deze studies nader bespreken.
Een belangrijk overzichtsartikel, is dat van Cohen (1981), vooral vanwege de hoeveelheid studies die in de analyse betrokken werden en de zorgvuldige manier van inventariseren en analyséren van de onderzoeksopzetten. Cohen (1981) verrichtte een meta-analyse op 68 'multisection validity studies' geselecteerd op de volgende drie criteria: -1) het onderzoek moest betrekking hebben op studentoordelen in de onderwijspraktijk (dus niet afkomstig zijn uit experimentele- of laboratoriumsituaties), -2) de eenheid van analyse moest de groep of klas zijn, en -3) de gegevens moesten gebaseerd zijn op een multisection kursus met op het einde een traditionele kennistoets die in alle secties werd afgenomen.
Uit de meta-analyse bleek dat over de 68 multisection kursussen de studieprestaties gemiddeld .50 gecorreleerd waren met studentoordelen betreffende doceervaardigheid van de docent ('skill'), .47 met het totaaloordeel over de kursus ('overall course rating'), .47 met de structuur/opzet van de kursus ('structure'), .47 met studievoortgang ('study progress'), en .43 met het totaal oordeel over de docent ('overall instructor rating'). Items die betrekking hadden op de moeilijkheidsgraad van de cursus, correleerden niet of negatief met studieprestaties. De correlaties waren hoger als de docenten een full-time aanstelling hadden, als studenten reeds de uitslag van hun toets kenden, of als de toets voorafging aan de invulling van de vragenlijst, en als de eindtoets niet door docenten van de secties was opgesteld maar door een docent die geen les had gegeven in de kursus. Cohen (1981) concludeerde op basis van de meta-analyse dat studentoordelen een valide maat zijn voor de kwaliteit van het gegeven onderwijs.

Feldman (1976) analyseerde eveneens een aantal onderzoeken waarin studentoordelen gecorreleerd werden met studieprestaties. Uit zijn analyse bleek dat als de eenheid van analyse werd gevormd door de individuele student, de correlaties tussen oordeel en prestaties voor het merendeel varieerden tussen .14 en .27. Als de groep of klas de eenheid van analyse vormde, lagen de correlaties iets hoger.
Aan het valideren van studentoordelen op basis van correlaties tussen oordelen en leerresultaten, kleven een aantal problemen. Cohen (1981) noemt de volgende vier problemen:1) de keuze van de eenheid van analyse, 2) de 'ability-rating bias', 3) de multidimensionaliteit van oordelen en 4) het design van het onderzoek. Marsh (1984) en Centra (1979) noemen nog een vijfde probleem namelijk de zogenaamde 5) 'grading leniency bias'. In het onderstaande gedeelte zal aan deze vijf problemen aandacht geschonken worden, omdat zij van wezenlijk belang zijn voor de interpretatie van de onderzoeksresultaten in hoofdstuk 4.

1) Eenheid van analyse.

Studentoordelen kunnen zowel op individueel niveau, als op groepsniveau met studieprestaties gecorreleerd worden. In het laatste geval worden groepsgemiddelden met elkaar in verband gebracht. Als individuele studenten de eenheid van analyse vormen, krijgt men antwoord op de vraag in hoeverre studenten die hoge leerresultaten behaald hebben, hogere beoordelingen aan docenten of kursussen hebben gegeven onafhankelijk van de klas of groep waarin studenten gezeten hebben. Als de groep of klas de eenheid van analyse vormt, krijgt men antwoord op de vraag in hoeverre beter beoordeelde docenten een bijdrage leverden aan betere leerresultaten van groepen of klassen. In overzichtsartikelen wordt er daarom op gewezen dat de groep of klas de eenheid van analyse moet vormen, indien men wil weten in hoeverre studentoordelen differentiëren tussen docenten die wel of geen bijdrage aan het leerproces geleverd hebben. (Cohen, 1981; Dowell & Neal, 1982; Marsh, 1984).

2) Ability rating bias (fouten als gevolg van verschil in voorkennis).

Het tweede probleem wordt omschreven als 'ability rating bias'. Deze vorm van bias ontstaat indien groepen of klassen voorafgaande aan de cursus al van elkaar verschillen wat betreft hun voorkennis of motivatie. Als het aantal groepen binnen een kursus groot is en de groepsgrootte klein, kan het voorkomen, ook al is er sprake van een aselecte toedeling van studenten aan groepen, dat er groepen ontstaan die wezenlijk verschillen van andere groepen wat betreft voorkennis, motivatie en studievaardigheden (in het Engels aangeduid als "abilities"). Verschillen in studieprestaties kunnen dan te wijten zijn aan verschillen tussen groepen die al bestonden voor het onderwijs een aanvang had genomen. Correlaties

tussen studentoordelen en studieprestaties kunnen dan aanleiding geven tot een over- of onderschatting van de werkelijke relatie die veroorzaakt wordt door toevallige verschillen in "ability". Vandaar dat men volgens Cohen (1981) voor een groot aantal klassen moet zorgen (meer dan 20), die ieder minimaal 15 studenten bevatten, aselect toegewezen aan die klassen.

3) Dimensionaliteit.

Het derde probleem heeft betrekking op multidimensionaliteit van studentoordelen. De meeste beoordelingsvragenlijsten (SEEQ, ICEQ, Purdue Rating Scale, Frey's Endeavor Instrument, SIRS) hebben een multidimensionele structuur (zie hoofdstuk 2). Vragenlijsten bevatten een aantal dimensies die kenmerken van onderwijs beschrijven (bijvoorbeeld, 'workload/difficulty', 'skill', 'rapport', 'enthusiasm', 'structure', 'interaction', 'feedback', 'evaluation' etc.). Per dimensie zijn een aantal items in de betreffende vragenlijst opgenomen.
Het construct 'kwaliteit van onderwijs' wordt op deze manier niet als een ééndimensionaal kenmerk gezien. Cohen (1981) en Marsh (1984) wijzen erop dat niet iedere dimensie van de vragenlijst (bijvoorbeeld moeilijkheidsgraad van de kursus) gecorreleerd hoeft te zijn met toetsresultaten van studenten. Vandaar dat een nul-correlatie in sommige gevallen niet altijd wijst op nonvaliditeit van studentoordelen. Op grond daarvan is het onverstandig de totale somscore van de verschillende dimensies te gebruiken om deze te correleren met toetsresultaten.

4) Onderzoeksdesign.

Het laatste probleem heeft betrekking op het onderzoeksdesign van multisection validiteitsstudies. De vraag is of bepaalde kenmerken van het design van invloed zijn op de correlaties tussen studentoordelen en toetsresultaten. Cohen (1981) noemt twintig factoren waarmee men het design van deze studies kan classificeren. Deze factoren hebben betrekking op het aantal klassen, de samenstelling van de klassen (aselecte toedeling versus vrije inschrijving), de opzet van de toets aan het einde van de kursus, de objectiviteit in scoring van de toets, de controle op voorkennis van studenten, het tijdstip van afname van de vragenlijsten, de duur van de kursus, de kwaliteit van de onderzochte kursus, het onderwerp van de kursus, de plaats van de kursus in het curriculum, de ervaring van de docent, etc. Uit de meta-analyse van Cohen (1981) blijkt dat slechts drie factoren de correlatie tussen studentoordelen met betrekking tot de dimensie 'overall instructor rating' en toetsresultaten beïnvloeden:
-1 het tijdstip van afname van de vragenlijsten
(voor of na de toets),

-2 de objectiviteit in scoring en opzet van de toets (de toets wordt gemaakt en nagekeken door een docent die geen les gaf in de kursus), en
-3 de ervaring van de docent (full-time aanstelling vs part-time).

Eén factor dient nog apart vermeld te worden namelijk het aantal secties waarop de studie betrekking heeft. Cohen (1981) rapporteert dat er een non-lineair verband bestaat tussen het aantal secties in de studie en de correlatie tussen studentoordeel en studieprestatie. Als het aantal klassen groter is dan twintig, is er sprake van een correlatie van .37 tussen 'overall instructor rating' en studieprestatie. Bij minder dan twintig klassen zijn de uitkomsten van de studies nogal variabel. Alle andere factoren beïnvloeden de correlatie niet.

Ter illustratie van de problemen die men kan tegenkomen bij multisection validatiestudies, zal een onderzoek beschreven worden dat opmerkelijke resultaten opleverde. Rodin en Rodin (1972) verrichtten een multisection validatie-onderzoek in een 'undergraduate calculus course' met 239 studenten. Alle studenten hadden drie dagen per week college van de voor de kursus verantwoordelijke docent. De overige twee dagen werden practica gegeven door onderwijsassistenten. In de practica werden sommen gemaakt, konden studenten vragen stellen over de leerstof en werden oude tentamenopgaven besproken. De bijeenkomsten waren vooral bedoeld voor studenten die extra hulp nodig hadden om de wiskundeproblemen op te lossen. Studenten waren vrij om gedurende de kursus van onderwijsassistent te wisselen. Bovendien waren de practicumbijeenkomsten niet verplicht. Er waren 12 secties en 6 onderwijsassistenten. Iedere onderwijsassistent gaf les aan twee secties. De kursusinhoud werd omschreven door 40 categorieën wiskundeproblemen. Over iedere categorie moesten studenten een toets afleggen. Als een student zakte, kon hij tot zes keer toe aan een herkansing deelnemen. De volgorde waarin de problemen opgelost moesten worden lag vast. De problemen werden nagekeken door de onderwijsassistent. Als een student een fout maakte bij de oplossing van het probleem, kreeg hij ongeacht het soort fout, de score nul voor dat probleem. Het eindcijfer voor de kursus werd bepaald door het aantal goed gemaakte problemen.

Aan het einde van de kursus vulden de studenten een vragenlijst in die betrekking had op de doceerkwaliteit van de onderwijsassistent. De verantwoordelijke docent werd niet beoordeeld. Als criterium voor studieprestatie gebruikten de onderzoekers het aantal goed opgeloste wiskundeproblemen. De correlatie tussen het gemiddelde studentenoordeel per sectie en studieprestatie bedroeg -.75 (N=12). Op basis van dit onderzoeksresultaat concludeerden Rodin en Rodin, dat naarmate studenten minder in een kursus leerden, docenten beter beoordeeld werden: 'students rate most highly instructors from whom they learn least'. Deze studie werd, gezien haar sensationele conclusie en publikatie in het tijdschrift

Science, de meest geciteerde validatiestudie op het gebied van studentoordelen (Cohen, 1983). Later is aangetoond, onder andere door Doyle (1975), dat in dit onderzoek een reeks van fouten was gemaakt.
De fouten hadden betrekking op de wijze van toetsing, de indeling van de secties, en datgene wat door de studentoordelen gemeten werd. De oordelen hadden geen betrekking op de kursus in zijn geheel, doch slechts op de onderwijsassistenten. De practicumbijeenkomsten waren niet verplicht, waardoor de goede studenten waarschijnlijk wegbleven. Bovendien konden studenten van onderwijsassistent wisselen. Daardoor was het onduidelijk op welke assistent, c.q sectie, het oordeel betrekking had. Doordat de secties niet random ingedeeld waren en grotendeels bestonden uit de zwakkere studenten, was het mogelijk dat zwakke studenten de assistent positief beoordeelden en goede studenten een neutraal oordeel gaven. Dit had tot gevolg dat een negatieve correlatie gevonden werd. Een ander probleem was het aantal secties. Bij een dergelijk aantal treden snel 'sampling errors' op waardoor de correlatie sterk beïnvloed kan worden. Een laatst probleem was dat men als maat voor het studentenoordeel de somscore van alle items van de vragenlijst hanteerde. Op deze wijze werd de doceerkwaliteit van de docent als een unidimensionale trek opgevat. In de literatuur is er steeds op gewezen dat dit een onhoudbare veronderstelling is.
Dit alles had tot gevolg dat er een negatieve correlatie gevonden werd tussen studentoordelen en toetsresultaten. Een dergelijk verband is in latere onderzoeken (door andere onderzoekers) niet meer gevonden. Des te opmerkelijker is het dat deze studie zoveel aandacht heeft gekregen.

5) Grading leniency bias.

"Grading leniency bias" verwijst naar de mogelijkheid dat studenten een docent beter beoordelen naarmate deze soepeler is in de manier van tentamineren en cijfers toekennen. De correlatie tussen studentoordelen en leerresultaten verwijst dan niet naar een werkelijk verband tussen de kwaliteit van onderwijs en leerresultaten, maar is het resultaat van contaminatie. Howard en Maxwell (1980) omschrijven de werking van 'grading leniency bias' als volgt: '.. easy graders receive better evaluations than hard graders *because* they are easy graders'. Aanhangers van de 'grading leniency bias' hypothese (bijvoorbeeld, Rodin & Rodin, 1972) stellen dat naarmate studenten hogere cijfers krijgen of verwachten, zij docenten een hogere beoordeling zullen geven, hetgeen resulteert in een hoge correlatie tussen studentoordelen en leerresultaten. De impliciete veronderstelling is dat als een docent studenten hoge cijfers geeft en weinig inspanning van studenten eist, die docent een positievere beoordeling krijgt. De docent wordt beloond voor dit gedrag in de vorm van een positieve beoordeling. Hij wordt beoordeeld op zijn 'bereidwilligheid' om makkelijke toetsen te geven of om soepele

normen bij de beoordeling te hanteren, in plaats van zijn manier van onderwijzen. Costin, Greenough, en Menges (1971), Page (1974), en Centra (1979) wijzen er in overzichtsartikelen op, dat er over het geheel genomen geen aanwijzingen zijn dat studenten zich in hun oordelen laten beïnvloeden door behaalde of verwachte toetscijfers. Uit onderzoek van Costin Greenough en Menges (1971) bleek dat het niet aannemelijk is dat er een relatie bestaat tussen verwacht of behaald tentamenresultaat en studentbeoordeling van het onderwijs. Deze relatie werd onderzocht door studentoordelen te correleren met studieprestaties, waarbij men controleerde voor de invloed van het verwachte of behaalde tentamenresultaat door de vragenlijsten voor of na de toets af te nemen. Slechts in een enkel onderzoek (bijvoorbeeld Holmes, 1972) vond men dat studenten een docent lagere beoordelingen gaven als hun studieprestatie lager was dan zij verwacht hadden. Marsh (1984) stelt dat er drie hypotheses mogelijk zijn voor het verklaren van positieve correlaties tussen studentoordelen en toetsresultaten:

-1 beter onderwijs leidt tot betere studieprestaties en betere studentbeoordelingen (de criteriumvaliditeitshypothese),
-2 betere toetsresultaten leiden tot meer satisfactie (dus hogere beoordeling), ongeacht het feitelijk gedrag van de docent of de kwaliteit van het onderwijs (grading leniency hypothese),
-3 verschillen in studentkenmerken tussen secties (voorkennis, motivatie of studievaardigheid) veroorzaken verschillen tussen docenten en toetsresultaten (ability rating hypothese).

Als de eerste hypothese waar zou zijn, dan pleit dit voor de validiteit van studentoordelen. De tweede hypothese verwijst naar het bestaan van grading leniency bias. De derde hypothese verwijst naar de invloed van beginkenmerken van studenten. Marsh (1984) kon slechts in twee studies (Marsh, Fleiner & Thomas, 1975; Marsh & Overall, 1980) de tweede hypothese verwerpen. Hij concludeert dat, mede op basis van de resultaten van Cohen's (1981) meta-analyse, het verwachte of behaalde studieprestatie waarschijnlijk altijd van invloed is op het studentoordeel. Het probleem blijft echter hoe groot die invloed is. Howard en Maxwell (1980) onderzochten, met behulp van pad-analytische technieken in hoeverre er sprake kan zijn van een causale relatie tussen verwacht of behaald studieprestatie en studentoordeel. Hun hypothese was: beter onderwijs leidt tot een hogere motivatie van studenten, een hogere motivatie leidt tot betere studieresultaten en tot een hogere satisfactie met betrekking tot het onderwijs (c.q. betere beoordeling). Zij onderzochten in hoeverre deze interpretatie van de grading leniency bias, correlaties tussen studieprestatie en studentoordelen kon verklaren. Uit hun onderzoek bleek dat hun hypothese niet verworpen kon worden. De resultaten wezen wel op een verband tussen motivatie, satisfactie en studieresultaat, maar studieresultaat was nauwelijks

gerelateerd aan satisfactie. Er was met andere woorden geen sprake van causale relatie tussen verwacht of behaald tentamenresultaat en studentoordeel. Zij concludeerden op basis van hun uitkomsten dat de relatie tussen studentoordelen en tentamenresultaat '...might be viewed as a welcomed result of important causal relationships among other variables rather than simply evidence of contamination due to grading leniency'.
Grading leniency bias is een vorm van bias waar in onderzoek veel aandacht aan is besteed. Het blijkt echter bijzonder moeilijk te zijn om aan te tonen dat studenten in hun oordeel beïnvloed worden door het verwachte of behaalde tentamenresultaat. Immers correlaties tussen studentoordelen en toetsresultaten kunnen zowel een bevestiging als een aantasting van de validiteit van studentoordelen betekenen. De vraag blijft daarom bestaan, of er daadwerkelijk sprake is van een causaal effect, en zo ja of dit effect de validiteit op een negatieve manier beïnvloedt.

3.7.3 Veldstudies: onderzoek naar de constructvaliditeit van studentoordelen.

De criteriumvaliditeitsbenadering levert, zoals uit het voorgaande bleek, gemengde resultaten op. Voorzover validiteitscoëfficiënten gevonden worden, zijn ze tamelijk laag. Vandaar dat onderzoekers in dit domein constructvalidatiebenaderingen gingen toepassen. Deze aanpak wordt vooral gerapporteerd vanaf 1979. De meest gebruikelijke benaderingen behelzen onderzoek naar de factorstructuren van oordelen van docenten en studenten op identieke vragenlijsten, of het analyseren van multitrait-multimethod matrices (Braskamp, Caulley, & Costin, 1979; Doyle & Crichton, 1978; Marsh, Overall & Kesler, 1979; Marsh, 1982; Marsh & Hocevar, 1984). Marsh, Overall en Kesler (1979) bijvoorbeeld, verzamelden in 207 kursussen docent- en studentoordelen met behulp van een identieke vragenlijst. De mediane correlatie tussen docent- en studentoordelen op de subschalen van de vragenlijst bedroeg .49. Factoranalyse van de vragenlijst liet dezelfde factoren bij studenten en docenten zien. De onderzoekers concludeerden dat er een redelijk hoge mate van overeenstemming bestaat tussen student- en docentoordelen over onderwijs. De resultaten zijn volgens hun een aanwijzing voor de validiteit van studentoordelen. Howard, Conway en Maxwell (1985) correleerden studentoordelen met docentoordelen, oordelen van getrainde observatoren, collega-docenten en oud-studenten. Uit hun onderzoek bleek dat er wel sprake was van overeenstemming tussen docenten en studenten, maar dat studenten en oud-studenten meer valide oordelen gaven dan de andere beoordelaars. Een mogelijke verklaring voor dit resultaat was dat studenten waker 'blootgesteld' worden aan de betreffende docenten dan de andere beoordelaars.
Doyle en Crichton (1978) vonden in hun onderzoek een mediane correlatie van .47 tussen studentoordelen en zelfbeoordelin-

gen van tien docenten die ieder in een sectie van een multi-section kursus les gaven.
Braskamp, Caulley en Costin (1979) onderzochten de relaties tussen studentoordelen, docentoordelen en studieprestaties. In een inleidende kursus psychologie, die twee keer per jaar gegeven werd en een semester besloeg, werd aan studenten en docenten (23 docenten in de najaarskursus, 17 docenten in de voorjaarskursus) een identieke vragenlijst over de kursus voorgelegd, die 24 items bevatte. De items waren gegroepeerd in vijf subschalen namelijk: 'student involvement' (betrokkenheid van student bij het onderwijs), 'teacher support' (docentbegeleiding), 'negative affect' (sfeer tijdens de kursus), 'teacher control' (sturing van het onderwijs), en 'teacher skill' (vaardigheden van de docent). De docenten waren student-assistenten die vergevorderd waren met hun studie. Ze gaven ieder aan twee secties les. Zeventien docenten gaven zowel in de najaars- als voorjaarskursus les. Met betrekking tot het najaarssemester werden slechts enkele significante correlaties tussen subschalen van de docenten- en studentenvragenlijst gevonden. De correlaties hadden voornamelijk betrekking op de subschaal 'teacher support'. Ze varieerden tussen .48 en .68. In de voorjaarskursus werden meer significante correlaties gevonden. Deze hadden betrekking op de subschalen 'student involvement', 'teacher support' en 'teacher skill'. Studenten en docenten werd tevens in beide kursussen gevraagd een globaal oordeel te geven over de kursusinhoud, de docent, en de kursus in zijn totaliteit. De onderzoekers rapporteren een correlatie van .39 tussen student- en docentoordeel in de najaarskursus, en een correlatie van .69 in de voorjaarskursus. De docentoordelen correleerden niet met de toetsresultaten van studenten. Van de studentoordelen correleerde, in de voorjaarskursus, alleen de 'teacher control' schaal met de toetsresultaten van studenten (correlatie: .58). De correlaties tussen docent- en studentoordelen wezen volgens de onderzoekers op de validiteit van studentoordelen. De correlatie tussen studentoordelen ('teacher skill' subschaal) en toetsresultaten was volgens hun in overeenstemming met eerder verricht onderzoek. Op basis van deze resultaten concludeerden zij dat docentoordelen een nuttig hulpmiddel kunnen zijn bij de evaluatie van onderwijs. Verschillen in oordelen tussen beide groepen waren volgens hun te wijten aan het verschil in perspectief op het onderwijs.

3.8 Conclusies.

In dit hoofdstuk is nader ingegaan op de problematiek rond de betrouwbaarheid en validiteit van studentoordelen.
Uit de in dit hoofdstuk beschreven onderzoeken blijkt, dat de betrouwbaarheid van studentoordelen over het geheel genomen hoog is. Onderzoek naar de validiteit van studentoordelen heeft over het algemeen redelijk goede resultaten opgeleverd. In laboratoriumstudies is onderzocht in hoeverre studenten

in hun oordeel beinvloed worden door hun algemene sympathie voor docenten. Uit deze studies bleek dat studenten in staat waren om onafhankelijk van persoonskenmerken (aardig-onaardig) van docenten, gedrag van docenten op een accurate manier te beschrijven. Studenten leken wel door persoonskenmerken van docenten beinvloed te worden wat betreft hun evaluatieve verwachtingen over de kwaliteit van de colleges van deze docenten.

Uit veldonderzoeken is veelvuldig gebleken dat studentoordelen als een valide indicator voor de kwaliteit van het onderwijs beschouwd kunnen worden. Onderzoeken die studentoordelen valideren aan externe criteria zoals studieprestaties, leveren over het geheel genomen aanvaardbare resultaten op. De constructvaliditeit van studentoordelen blijkt eveneens redelijk tot goed te zijn. Samenvattend kan gesteld worden dat de in dit hoofdstuk gerapporteerde bevindingen de conclusie rechtvaardigen dat studentoordelen gebruikt kunnen worden voor het beoordelen van de kwaliteit van het onderwijs.

HOOFDSTUK 4. BETROUWBAARHEID EN VALIDITEIT VAN STUDENTOORDELEN: EMPIRISCHE STUDIES.

4.1 Inleiding.

De inhoud van de items die opgenomen zijn in een beoordelingsvragenlijst, is bepalend voor de kwaliteit van de informatie die verkregen wordt met behulp van studentoordelen. Vandaar dat in dit hoofdstuk uitvoerig aandacht besteed wordt aan de vraag in hoeverre de beoordelingsvragenlijst die binnen het evaluatiesysteem van de medische faculteit (RL) gebruikt wordt, voldoet aan de eisen die men kan stellen aan de betrouwbaarheid en validiteit van dit type instrument.
In dit hoofdstuk worden zes empirische studies beschreven naar de betrouwbaarheid en validiteit van de beoordelingsvragenlijst voor studenten.
In de eerste studie wordt met behulp van principale componentenanalyse onderzocht of de gebruikte vragenlijst een multidimensionele struktuur heeft. In de tweede studie wordt de interne consistentie onderzocht van schalen die geconstrueerd zijn op basis van de uitkomsten uit studie 1. In studie 3 wordt onderzoek beschreven naar de interbeoordelaars-betrouwbaarheid van de studentoordelen. Daarna wordt in studie 4 een onderzoek beschreven over de criteriumvaliditeit van studentoordelen. Tenslotte worden in de studies 5 en 6 onderzoeken naar de construct-validiteit van studentoordelen beschreven.

4.2 Instrumenten, onderzoekspopulatie en procedure.

Gedurende de academiejaren 1981/1982, 1982/1983, 1983/1984 en 1984/1985 werden, met behulp van de in hoofdstuk 2 beschreven vragenlijst, studentoordelen verzameld die betrekking hadden op de blokken in de eerste vier jaren van het onderwijsprogramma van de medische faculteit. In het onderzoek waren 74 blokken betrokken. De eerste blokken uit het academiejaar 1981/1982, namelijk blok 1.1, 2.1, 3.1 en 4.1 (blokken zijn genummerd, waarbij het eerste getal het studiejaar aangeeft en het tweede getal de plaats van het blok binnen het studiejaar), maakten geen deel uit van het onderzoek. De vragenlijst die in deze blokken gebruikt werd, week wat betreft zijn inhoud teveel af van de daarna gebruikte lijsten.
Tevens werden in het academiejaar 1984/1985 in ieder blok tutoroordelen verzameld met behulp van een vragenlijst die was gebaseerd op de vragenlijst die aan studenten werd voorgelegd. De tutorvragenlijst had 32 items gemeenschappelijk met de studentvragenlijst. De gegevens van deze tutorvragenlijst zullen gebruikt worden in onderzoek naar de studentoordelen. Tutoroordelen uit de voorgaande academiejaren zijn in deze studies niet gebruikt. De in deze periode gehanteerde vragenlijst bevatte namelijk te weinig gemeenschappelijke items met de studentvragenlijst, om als basis voor het verderop beschreven validatie-onderzoek te dienen.

Uit het voorgaande blijkt dat de in het onderzoek gebruikte vragenlijsten voor studenten en tutoren in de opeenvolgende academiejaren een aantal wijzigingen hebben ondergaan: nieuwe items zijn opgenomen, oude geschrapt of in formulering gewijzigd. In bijlage 1 zijn de veranderingen in de vragenlijsten in de verschillende academiejaren weergegeven. Tevens is aangegeven welke items uiteindelijk voor het onderzoek gebruikt zijn. In het onderzoek zijn alleen die items betrokken die binnen de gekozen onderzoeksperiode niet aan veranderingen onderhevig waren.
Alle analyses in de studies naar de betrouwbaarheid en validiteit van de studentoordelen zijn, met uitzondering van de interrater analyses, gebaseerd op studentoordelen die op het niveau van de onderwijsgroep werden geaggregeerd: per item werd de gemiddelde score van de studentoordelen binnen een onderwijsgroep berekend. Zoals in hoofdstuk 3 reeds werd opgemerkt, is voor de meeste analyses het gemiddelde groepsoordeel te prefereren boven individuele oordelen. Immers individuele oordelen komen niet onafhankelijk van elkaar tot stand. Statistisch gezien zijn de individuele studentoordelen 'genest' binnen onderwijsgroepen. Een blok bestaat immers uit een aantal verschillende onderwijsgroepen (ieder met hun eigen samenstelling en tutor) met ieder 8 tot 10 studenten. Dit betekent dat in een aantal gevallen verschillende groepen studenten, verschillende objecten beoordelen (bijvoorbeeld het functioneren van de onderwijsgroep en de tutor) en dat in andere gevallen verschillende groepen hetzelfde object beoordelen (bijvoorbeeld de taken, de bloktoets). De gegevens over de 1089 onderwijsgroepen zijn gebaseerd op 7879 individuele oordelen, een gemiddelde van 7.2 oordelen per onderwijsgroep. Het responspercentage bedroeg 94%. In tabel 4.1 wordt een overzicht gegeven van het aantal onderwijsgroepen dat in het onderzoek betrokken was.

Tabel 4.1: Aantal onderwijsgroepen dat per academiejaar in het onderzoek betrokken was.

Academiejaar	Aantal Onderwijsgroepen
1981/1982	173
1982/1983	263
1983/1984	296
1984/1985	357
Totaal:	1089

In enkele studies zijn ook bloktoetsgegevens gebruikt. Zoals in hoofdstuk 2 reeds werd aangegeven, worden bloktoetsen door de planningsgroepen van het betreffende blok samengesteld.

De toetsen hebben het karakter van een klassieke studietoets. Ze omvatten gemiddeld 200 items van het type juist/onjuist. De resultaten voor de bloktoets worden uitgedrukt in een procentuele goed- en goed-min-fout score. De procentuele goed-score wordt berekend door het aantal goed gemaakte items te delen door het totaal aantal items en vervolgens te vermenigvuldigen met 100. De goed-min-fout score wordt op analoge wijze berekend.
Alle analyses werden verricht met het statistische pakket SPSSX (Nie, Hull, Jenkins, Steinbrenner & Bent, 1983) m.u.v. studie 6 waarvoor het pakket LISREL IV werd gebruikt (Jöreskog & Sörbom, 1978).

4.3 Studie 1: Onderzoek naar de multidimensionaliteit van de beoordelingsvragenlijst voor studenten.

4.3.1 Inleiding.

In hoofdstuk 1 werd gesuggereerd dat "onderwijskwaliteit" wellicht geen unidimensioneel construct is en dat het daarom noodzakelijk zou kunnen zijn gebruik te maken van multidimensionele vragenlijsten die verschillende facetten van dit construct te meten. In hoofdstuk 2 werd uitvoerig aandacht besteed aan de theoretische en praktische achtergronden op grond waarvan de hier onderzochte beoordelingsvragenlijst geconstrueerd werd. Daaruit bleek dat een gecombineerde constructiemethode (inductie en deductie) gebruikt is om de vragenlijst te construeren. Voorts werd in hoofdstuk 3 op basis van verschillende onderzoeken geconcludeerd, dat beoordelingsvragenlijsten een multidimensionele structuur bezitten mits voldoende aandacht besteed is aan de constructie daarvan (Marsh, 1984).
In hoofdstuk 2 werd geschreven dat aan de beoordelingsvragenlijst voor studenten, theorieën over onderwijs in het algemeen en theorieën over probleemgestuurd onderwijs in het bijzonder ten grondslag liggen. Bovendien is bij de constructie van de lijst rekening gehouden met opinies van studenten en docenten over kenmerken van kwalitatief goed probleemgestuurd onderwijs. In de vragenlijst zijn op basis van deze uitgangspunten groepen onderling samenhangende items geformuleerd, die als operationalisatie van die theorieën en opinies gezien kunnen worden. Deze itemgroepen (betreffende onderwijsgroep, tutor, blokboek, etc.) kunnen als dimensies beschouwd worden die de onderliggende structuur van de vragenlijst vormen.
De vraag die zich in deze studie voordoet is of de theoretische structuur van de vragenlijst die binnen de medische faculteit gebruikt wordt, overeenkomt met de empirische structuur. De mogelijkheid bestaat immers dat aan studentoordelen andere beoordelingsdimensies ten grondslag liggen dan in de vragenlijst verondersteld wordt. Bovendien bestaat de mogelijkheid dat studenten in hun oordeel beïnvloed worden

door een aantal factoren die feitelijk niet gerelateerd zijn aan de kwaliteit van het genoten onderwijs. In hoofdstuk 3 werden twee groepen factoren genoemd die het oordeel van studenten kunnen beïnvloeden: contingent error en noncontingent error (paragraaf 3.2). Beide meetfoutbronnen zijn van invloed op de hoogte van de correlaties tussen de geobserveerde variabelen en beïnvloeden daarbij tegelijkertijd de onderliggende empirische dimensionale structuur van de beoordelingsvragenlijst.
In de nu volgende studie wordt met behulp van een factoranalytische techniek, de principale componentenanalyse, onderzocht of de theoretisch veronderstelde multidimensionele structuur van de in hoofdstuk 2 beschreven vragenlijst, overeenkomt met de empirische structuur. Bovendien wordt aandacht geschonken aan de vraag of halo-effecten of effecten van impliciete theorieën, een aannemelijke verklaring vormen voor de gevonden resultaten.

4.3.2 Methode.

In deze studie werden uitsluitend de gegevens van de onderwijsgroepen uit het academiejaar 1984/1985 gebruikt om de vragenlijst te analyseren. Een principale componenten analyse werd met behulp van het statistische pakket SPSSX (Nie, Hull, Jenkins, Steinbrenner & Bent, 1983) uitgevoerd op een databestand dat alle onderwijsgroepen uit het academiejaar 1984/1985 omvatte. De datamatrix bestond uit 357 onderwijsgroepen maal zestig items. De itemscores waren gemiddelde scores die per onderwijsgroep berekend werden. De principale componenten werden geëxtraheerd op basis van het Kaiser-criterium (eigenwaarde van de componenten groter of gelijk aan 1). Na de componentenextractie volgde een oblique rotatie.

4.3.3 Resultaten.

In tabel 4.2, 4.3 en 4.4 worden de resultaten van de principale componenten analyse weergegeven. Tabel 4.2 bevat resultaten van de extractie van de principale componenten, voordat de rotatie plaatsvond. Tabel 4.3 bevat de resultaten na de oblimin-rotatie van de principale componenten, namelijk de factor-patroon matrix. Tabel 4.4 geeft de correlaties tussen de geroteerde principale componenten weer.
Het Kaiser-criterium werd toegepast om het aantal componenten te bepalen dat gebruikt wordt voor de verdere analyse. Volgens dat criterium worden factoren of principale componenten niet in de verdere analyse opgenomen als deze minder variantie verklaren dan willekeurige andere variabelen waarop de analyse verricht wordt. Kim en Mueller (1978b) noemen als tweede veelgebruikt criterium de Scree-test van Cattel (1966). Dit is een test waarin eerst een grafiek opgesteld wordt van de successieve eigenwaarden. Daarna wordt gekeken op welke punten een relatief sterke afname in de eigenwaar-

den te zien is. In de verdere analyse wordt dan uitgegaan van het aantal componenten dat aan het betrokken punt vooraf ging. Dit criterium geeft echter onvoldoende uitsluitsel over het exacte aantal componenten dat geëxtraheerd moet worden. Het is daarom in deze studie meer een hulpmiddel bij het interpreteren van de resultaten.
In tabel 4.2 worden de eigenwaarden van de principale componenten weergegeven, de percentages verklaarde variantie en het cumulatieve percentage verklaarde variantie.
Voor de interpretatie van de resultaten in deze tabel zijn twee criteria van belang: 1) het aantal componenten dat gevonden wordt in verhouding tot het aantal variabelen dat in de analyse betrokken is, en 2) de hoogte van de successieve eigenwaarden. Saal, Downey en Lahey (1980) wijzen erop dat de invloed van halo-effecten en impliciete theorieën, op twee manieren bij de extractie van componenten in een principale componentenanalyse merkbaar kan worden. In de eerste plaats is het aantal principale componenten afhankelijk van de beïnvloeding van studentoordelen door halo-effecten en impliciete theorieën: hoe minder componenten, hoe meer kans op het optreden van deze effecten. In de tweede plaats kunnen dergelijke effecten zichtbaar worden in de verhouding verklaarde variantie van de eerste component en de resterende componenten: als de eerste component naar verhouding zeer veel variantie verklaart ten opzichte van andere componenten, kan er sprake zijn van bovengenoemde effecten.
Uit de resultaten in tabel 4.2 blijkt dat in de analyse 15 principale componenten geëxtraheerd werden met een eigenwaarde groter dan 1. In totaal verklaren de componenten 71 procent van de variantie. Gegeven de zestig items die in de analyse betrokken waren is het aantal componenten relatief groot. Uit de hoogte van de successieve eigenwaarden blijkt dat de eerste component in verhouding tot de andere componenten, niet teveel variantie verklaart. In vergelijking met in de literatuur beschreven resultaten van andere vragenlijsten, die als betrouwbaar en valide beschouwd worden, zijn de resultaten in tabel 4.2 zeer goed (Frey, Leonard & Beatty, 1975; Marsh, Overall & Kesler, 1979; Marsh, 1982).

Tabel 4.2 Eigenwaarden en percentages verklaarde variantie van de componenten.

Component	Eigenwaarde	Pct variantie	Cum pct
1	14.3	23.8	23.8
2	6.5	10.8	34.6
3	3.6	6.0	40.6
4	2.7	4.6	45.2
5	2.5	4.1	49.3
6	2.1	3.5	52.8
7	1.8	3.0	55.8
8	1.6	2.7	58.4
9	1.5	2.5	61.0
10	1.5	2.5	63.5
11	1.4	2.3	65.8
12	1.3	2.2	68.0
13	1.2	1.9	69.9
14	1.1	1.8	71.7
15	1.0	1.7	73.4

Deze vijftien componenten werden vervolgens voor de oblique rotatie gebruikt. In tabel 4.3 worden de resultaten na de oblimin-rotatie weergeven. De tabel bevat de factorpatroon matrix waarin de ladingen van de items op de geroteerde principale componenten zijn opgenomen. In de tabel zijn alleen itemladingen op de geroteerde principale componenten weergegeven die groter of gelijk zijn aan .30, of kleiner dan of gelijk aan -.30.
Voor de interpretatie van de resultaten in de factorpatroon matrix zijn twee criteria van belang: 1) de hoogte van de itemladingen en 2) de structuur van de itemladingen. Het eerste criterium eist dat op een component een aantal variabelen een hoge lading (groter dan .70) heeft. Voor een eenduidige interpretatie van een component is het namelijk belangrijk dat een of meerdere variabelen hoog laden (Tabachnick & Fidell, 1983). Het tweede criterium vereist dat het merendeel van de variabelen slechts op één component een substantiële lading heeft (lading groter dan .30 of kleiner dan -.30). Naarmate de variabelen in de factorpatroonmatrix beter aan beide criteria voldoen is de matrix beter interpreteerbaar.

Tabel 4.3 Factorpatroonmatrix van de principale-componentenanalyse op de vragenlijst voor de beoordeling van het onderwijs in de Geneeskunde (N = 357 onderwijsgroepen).

		F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
1.	Taken															
V12	Taken leenden zich voor systematische aanpak	90														
V11	Taken waren duidelijk omschreven	79														
V10	Informatie blokboek was duidelijk	72														
V17	Taken geven aanleiding tot formuleren van leerdoelen	72														
V14	Taken gaven aanleiding tot zinnige groepsdiscussie	59		-33												
V15	Taken gaven voldoende aanleiding tot zelfstudie	49				38										
V22	Het was mogelijk blokdoelstellingen te realiseren	48														
V03	De doelstellingen waren duidelijk	36														
2.	Tutor															
V34	Leek op de hoogte van onderwijskundige uitgangspunten		90													
V44	Funktioneerde goed in zijn/haar rol		89													
V35	Gaf de indruk zijn/haar rol plezierig te vinden		83													
V33	Bezit goed begrip van doelstellingen van blok		82													
V37	Stelde regelmatig discussiestimulerende vragen		69													
V36	Stimuleerde tot hard werken		52													
V43	Evalueerde regelmatig gang van zaken		46									37		-40		31
3.	Onderwijsgroep															
V30	Iedereen leverde een actieve bijdrage			-84												
V29	De bijeenkomsten waren productief			-83												
V28	De bijeenkomsten waren prettig			-81												
V27	Iedereen hield zich aan de afspraken			-81												
V25	Het betekende een stimulans voor zelfstudieaktiviteiten			-81												
V26	Duidelijke afspraken m.b.t. te bestuderen stof			-66												
V01	Afgelopen periode prettig gewerkt			-66						33						
V24	Systematische werkprocedures bij aanpakken taken			-56												
V31	Een rem in de voortgang van studie			38												
4.	Skillslab.															
V45	Training lichamelijk onderzoek zinvol				79											
V46	Training therapeutische vaardigheden zinvol				78											
V48	Tevredenheid begeleiding bovengenoemde trainingen				78											
V47	Training laboratoriumvaardigheden zinvol				73											
V50	Simulatiepatiënten-kontakten zinvol				60							32				
5.	Moeilijkheidsgraad/Zwaarte															
V07	De leerstof van dit blok was moeilijk					86										
V04	Het programma vergde veel studietijd					77										
V59	Tijd in uren per week besteed aan literatuurstudie					66										
V21	Aanleiding onderwerpen uit basisvakken te bestuderen					53										
V05	Onderwerpen nuttig i.v.m. medische studie					47										
V08	Ik heb in dit blok veel opgestoken			-32		45			31							
V09	Aangeboden leerstof interessant					32			30							
6.	Bloktoets															
V57	Onderwerpen uit doelstellingen toetsen						82									
V56	Sloot aan op de bestudeerde onderwerpen						79									
V58	Zelfevaluatiemiddelen sloten aan op de inhoud						78									

		F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
7.	Leermiddelen															
V53	Voldoende AV-middelen beschikbaar							89								
V20	Grote variëteit aan hulpmiddelen bij taken							78								
V54	Beschikbare AV-middelen inhoudelijk goede kwaliteit							72								
8.	Algemeen oordeel blok															
V52	Voldoende verscheidenheid aan literatuur Studielandschap							41	-49							
V60	Cijfer van 1 t/m 10 voor het totale blok			-44					47							
9.	Voorkennis															
V06	Attitude t.a.v. gezondheidszorg beïnvloed									-77						
V02	Blok sloot goed aan op voorkennis								36	44						
10.	Inhoudsdeskundigen															
V55	Aantal malen inhoudsdeskundige geraadpleegd										-90					
V41	Tutor stimuleerde raadplegen van inhoudsdeskundigen										-86					
V42	Tutor stimuleerde gebruik leer- en evaluatiemiddelen										-55					
11.	Skillslab (sociale vaardigheden)															
V51	Tevredenheid over begeleiding bij trainingen											77				
V49	Training sociale vaardigheden zinvol			36								67				
12.	Studie onafhankelijk van blok															
V16	Onafhankelijk van blokboek gestudeerd												83			
V32	Onafhankelijk van leerdoelen onderwijsgroep gestudeerd												62			
13.	Sturing ondwg															
V38	Tutor stuurde met eigen vakkennis de discussie													76		
V13	Taken waren sterk gestructureerd													57		
14.	Variëteit															
V18	Grote variëteit aan onderwerpen in blokboek														-82	
V19	Grote variëteit aan taken in blokboek														-76	
15.	Tutor hard werken															
V40	Tutor controleerde gemaakte afspraken															73
V23	Hoeveelheid taken in blokboek moet worden uitgebreid															65
V39	Tutor stimuleerde het maken van afspraken			44												47

Uit de resultaten in tabel 4.2 blijkt dat het merendeel van de items aan beide criteria voldoet. Bijna iedere component bestaat uit een aantal items met hoge ladingen (groter dan .70). Component 14 vormt een uitzondering op dit gegeven. Bovendien blijken de meeste items slechts op een één component te laden. De structuur van de factorpatroonmatrix komt sterk overeen met de onderdelen van de vragenlijst. Zeven onderdelen van de vragenlijst (namelijk: algemene indruk, blokboek, onderwijsgroep, tutor, leermiddelen, skillslab, bloktoets) zijn in de factorpatroon matrix terug te vinden als geroteerde principale component.
De componenten hebben achtereenvolgens betrekking op:
1) beoordeling van de taken;
2) functioneren van de tutor;
3) functioneren van de onderwijsgroep;
4) kwaliteit skillslab-trainingen;
5) behandelde leerstof; 6) bloktoets;
7) leermiddelen;
8) algemeen oordeel over de kwaliteit van het blok;
9) voorkennis studenten;
10) inhoudsdeskundigen;
11) sociale vaardigheidstrainingen;
12) studie onafhankelijk van het blok;
13) sturing van de studie-aktiviteiten van de onderwijsgroep;
14) afwisseling in het aanbod van de taken in het blokboek;
15) stimulering door de tutor tot hard werken.
In tabel 4.4 zijn de correlaties tussen de geroteerde componenten weergegeven. Voor de interpretatie van de gegevens in deze tabel zijn twee criteria van belang: 1) de hoogte van de correlaties tussen de componenten, en 2) de hoogte van de correlaties tussen de eerste component en de andere componenten. Beide criteria geven inzicht in de vraag of halo-effecten en impliciete theorieën ten grondslag liggen aan de geïdentificeerde beoordelingsdimensies van studenten. Hoge correlaties tussen de componenten indiceren dat de dimensies sterk met elkaar samenhangen; sterker dan men in theorie zou mogen verwachten. Hoge correlaties tussen de eerste component en de andere componenten, wijzen op de mogelijkheid dat studenten in hun oordeel sterk beïnvloed worden door een enkele factor die sterk overheerste is bij de beoordeling van het onderwijs. Uit de resultaten in tabel 4.4 blijkt dat de correlaties over het geheel genomen laag zijn. De meeste correlaties liggen tussen -.30 en .30. Uit de resultaten blijkt eveneens dat er geen sprake is van sterke correlaties tussen de eerste componenten en de andere componenten.

Tabel 4.4 Correlatiematrix van de geroteerde principale componenten.

Component	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 Taken															
2 Tutor	21														
3 Onderwijsgr.	-29	-25													
4 Skillslab	21	-05	-14												
5 Zwaarte	20	01	-13	20											
6 Bloktoets	24	06	-17	26	16										
7 Leermiddl.	19	05	-18	25	21	28									
8 Alg.oordl.	10	-03	-06	12	10	12	11								
9 Voorkennis	03	-03	-02	03	05	03	01	02							
10 Inhoudsdesk	-06	-30	18	01	-02	02	-01	01	04						
11 Sociale Vaard.	04	12	00	16	12	01	08	-03	02	-04					
12 Onafh.studeren	-07	-13	15	03	01	-06	05	01	05	06	04				
13 Structuring	-03	09	03	-04	11	01	01	07	-01	-06	-03	01			
14 Afwisseling	-30	-18	16	-17	-19	-26	-22	-06	02	09	-08	06	01		
15 Tutor hard werken	01	24	-05	-07	02	-01	08	05	-05	-14	07	-03	07	02	

Noot. Correlaties zijn met 100 vermenigvuldigd.

4.3.4 Discussie en conclusies.

Inter-itemcorrelaties vormen de basis voor factoranalyse en principale componenten analyse. Deze correlaties kunnen beïnvloed worden door het optreden van systematische meetfouten (halo-effecten, impliciete theorieën). Dergelijke effecten beinvloeden het aantal factoren die in de analyse gevonden worden. De correlaties tussen de items zullen immers groter worden naarmate er meer sprake is van systematische meetfouten: het aantal factoren wordt dan automatisch kleiner. De correlatiematrix bevat in een dergelijk geval groepen items die niet discrimineren tussen verschillende dimensies van het gegeven onderwijs. Blijkt dat het aantal verschillende dimensies gering is, dan kan dit dus te wijten zijn aan systematische meetfouten zoals halo-effecten, of effecten van impliciete theorieën over het functioneren van de werkelijkheid. Hoe minder onderliggende dimensies gevonden worden, hoe groter de kans op het optreden van systematische meetfouten.
Uit de resultaten van de principale componentenanalyse blijkt dat: 1) naar verhouding een groot aantal componenten gevonden wordt, 2) dat deze componenten over het geheel genomen goed interpreteerbaar zijn, 3) dat deze grotendeels overeenkomen met de afzonderlijke onderdelen van de vragenlijst, en 4) dat de componenten onafhankelijk zijn. Deze uitkomsten lijken de conclusie te rechtvaardigen dat de onderzochte vragenlijst een multidimensionele structuur heeft waarmee verschillende facetten van het construct onderwijskwaliteit gemeten kunnen worden. Uit de resultaten van studie 1 kan bovendien gecon-

cludeerd worden dat er nauwelijks aanwijzingen zijn voor het optreden van halo-effecten of effecten van impliciete theorieën. Twee onderzoeksuitkomsten geven aanleiding tot deze conclusies.
In de eerste plaats blijkt uit de componentenextractie dat het aantal componenten en de corresponderende eigenwaarden niet op het optreden van halo-effecten of impliciete theorieën wijzen. Immers, een aantal van vijftien componenten binnen een vragenlijst met zestig items, sluit dergelijke effecten uit.
In de tweede plaats blijkt dat het optreden van halo-effecten in de vorm van een 'general factor' (bijvoorbeeld de aardigheid van de tutor, of de moeilijkheidsgraad van de leerstof) niet aannemelijk is. Een dergelijk effect zou immers tot uiting moeten komen in de proportie verklaarde variantie van de eerste component. De eerste component zou dan relatief veel variantie moeten verklaren en de volgende componenten relatief weinig.
Uit de resultaten blijkt dat de eerste component 22.9 procent van de totale variantie verklaart. De andere componenten verklaren naar verhouding minder variantie. De vraag is of dit een aanwijzing is voor het optreden van halo-effecten. Harde criteria voor het beantwoorden van deze vraag ontbreken echter. Wel kunnen de in deze studie gevonden resultaten vergeleken worden met resultaten uit ander onderzoek. Uit onderzoek van Marsh (1982a, 1982b) blijkt bijvoorbeeld dat in een vergelijkbare beoordelingsvragenlijst (de SEEQ) negen componenten gevonden worden op een aantal van 35 items. In totaal wordt 88% van de variantie verklaard door deze componenten. De eerste component verklaarde 56.8% van de variantie, de tweede 9.4%, de derde 6.7%, de vierde 4.3%, de vijfde 3.4%, de zesde 2.6%, de zevende 2%, de achtste 1.7%, en de negende 1.4%. De eigenwaarden bedroegen respectievelijk 19.9, 3.3, 2.3, 1.5, 1.2, 0.9, 0.7, 0.6, en 0.5. Volgens Marsh, die als expert op dit gebied beschouwd wordt,. worden studentoordelen zoals gemeten met de SEEQ, nauwelijks beïnvloed door halo. Men kan bij deze conclusie, gegeven de gerapporteerde resultaten, uiteraard een aantal vraagtekens plaatsen. Met betrekking tot de in studie 1 gerapporteerde resultaten kan echter geconcludeerd worden dat, in vergelijking met andere in de literatuur beschreven onderzoeken, de vragenlijst een multidimensionele structuur heeft die niet of nauwelijks beïnvloed wordt door systematische meetfouten. Onze conclusie is dan ook dat als systematische meetfouten een rol gespeeld hebben bij het tot stand komen van de oordelen van studenten, die rol klein moet. Dat blijkt uit het feit dat het percentage door de eerste component verklaarde variantie, vergeleken met ander onderzoek in dit domein, vrij gering is en uit het feit dat de componenten nauwelijks aan elkaar gecorreleerd zijn.

4.4 Studie 2: Schaalconstructie en betrouwbaarheid.

4.4.1 Inleiding.

In studie 1 is aangetoond dat de vragenlijst een multidimensionele structuur bezit en uit vijftien afzonderlijke dimensies bestaat. In deze studie zullen de resultaten uit de voorgaande studie als basis dienen voor de constructie van een aantal schalen waarmee kenmerken van probleemgestuurd onderwijs gemeten kunnen worden. Deze schalen dienen dus als operationalisatie van een aantal facetten die samen het construct 'de kwaliteit van probleemgestuurd onderwijs' omvatten. Met behulp van deze schalen kan in een latere studie (studie 4) de samenhang gemeten worden met andere indicatoren voor dit construct, bijvoorbeeld toetsprestaties. Met behulp van itemanalyse wordt de interne consistentie (alpha-betrouwbaarheid) bepaald van groepen items die samen een dimensie of facet vormen. Uit onderzoek van Feldman (1977) en Marsh (1984) is gebleken dat de alpha-betrouwbaarheid van dergelijke schalen meestal zeer hoog is.

4.4.2 Methode

Kim en Mueller (1978b) beschrijven niet minder dan zes procedures om op basis van factor-analyse of principale-componentenanalyse schalen te construeren: 1) de regressie-benadering, 2) het least squares criterium, 3) de Bartlett methode, 4) de orthogonaliteitsrestrictie, 5) de sommatie van variabelen met hoge factorladingen, en 6) de principale componentenschaling. De eerste vier procedures streven naar een maximale correspondentie tussen de factoren uit de factoranalyse en de geconstrueerde schalen. Afwijkingen tussen de gereproduceerde correlatiematrixen van respectievelijk de factoren en de geconstrueerde schalen worden met behulp van deze procedures geminimaliseerd. Deze procedures worden vooral gebruikt in situaties waarin meetinstrumenten aan factoranalyse onderworpen zijn. Procedure 5) en 6) worden vooral gebruikt bij meetinstrumenten die met behulp van principale-componentenanalyse onderzocht zijn. De laatste twee procedures berusten op het principe van sommatie van een aantal variabelen afkomstig uit een geroteerde principale component, waarbij aan die variabelen een bepaald gewicht wordt toegekend. In deze studie is methode 5) toegepast. Voor iedere component werden de items geselecteerd die samen een schaal vormen. Ieder item kreeg hetzelfde gewicht, namelijk: '1'. Itemselectie vond plaats aan de hand van de volgende criteria:

- de items moesten binnen een component een factorlading hebben die groter of gelijk was aan de absolute waarde van .30,
- de items moesten inhoudelijk aansluiten op het merendeel van de items binnen de component,

- items die op meerdere factoren laadden, werden in de schaal opgenomen op welke onderliggende component zij de grootste lading hadden.

Items die binnen een principale component een tegengestelde lading hadden, werden gehercodeerd (wisseling van de antwoordcategorieën). Per schaal werd de interne consistentie bepaald door coëfficiënt alpha te berekenen. Deze coëfficiënt geeft inzicht in de mate waarin items binnen een schaal, aspecten of facetten van een unidimensionaal construct meten.

Schalen werden geconstrueerd op basis van de in studie 1 beschreven principale-componentenanalyse. De vragenlijst die in deze studie gebruikt werd, was afkomstig uit het academiejaar 1984/1985. Dit impliceert dat sommige items uit deze vragenlijst niet zullen voorkomen in de vragenlijsten van voorgaande academiejaren (zie bijlage 1). Sommige schalen bevatten dus voor eerdere academiejaren missing items. Met behulp van SPSSX werd coëfficiënt alpha van de aldus verkregen schalen berekend. Gemiddelde itemscores van onderwijsgroepen vormden de eenheid van analyse.

4.4.3 Resultaten

In tabel 4.5 zijn de alpha's van de hierboven beschreven schalen per academiejaar weergegeven. Tevens is per academiejaar opgenomen het aantal items dat deel uitmaakt van de schalen. Uit de gegevens blijkt dat de alpha-betrouwbaarheden van de meeste schalen hoog zijn (circa .80). Dat betekent dat op deze schalen de items een redelijk homogene groep vormen en dat ze als afspiegeling gezien kunnen worden van de onderliggende dimensie uit de vragenlijst. Schaal 13 (structuur) heeft een zeer lage alpha-betrouwbaarheid.

Tabel 4.5: Alpha-betrouwbaarheden van de schalen.

	Academiejaar			
Dimensie	1981/1982	1982/1983	1983/1984	1984/1985
1 Taken	.89 (6)	.92 (8)	.91 (8)	.89 (8)
2 Tutor	.88 (7)	.86 (7)	.90 (7)	.90 (7)
3 Onderwijsgr.	.89 (6)	.94 (8)	.94 (8)	.94 (8)
4 Skillslab	--	--	.75 (5)	.84 (5)
5 Zwaarte	-- (1)	.81 (5)	.74 (5)	.84 (6)
6 Bloktoets	--	.71 (3)	.60 (3)	.78 (3)
7 Leermidd.	.67 (2)	.78 (4)	.79 (4)	.77 (4)
8 Globaal oord.	-- (1)	-- (1)	-- (1)	-- (1)
10 Inhoudsdesk.	--	--	.73 (3)	.74 (3)
11 Soc.Vaardigh.	-- (0)	-- (0)	.76 (2)	.81 (2)
12 Onafhank.	-- (1)	-- (1)	-- (1)	.53 (2)
13 Structuur	-- (1)	-.05 (2)	-.09 (2)	.25 (2)
14 Afwisseling	--	.82 (2)	.79 (2)	.81 (2)
15 Tutor hw	.85 (2)	.68 (2)	.78 (2)	.81 (2)
N	164	238	240	337

Tussen de haakjes zijn de aantallen items vermeld.

4.4.4 Discussie en conclusies.

Analyse van de items die deel uitmaken van de onderliggende dimensies van de vragenlijst, toont aan dat het merendeel van de schalen uit homogene groepen items bestaat. Eén schaal, namelijk schaal 13, heeft een opvallend lage alpha-betrouwbaarheid. Een verklaring hiervoor kan gelegen zijn in het optreden van weinig itemvariantie op de betreffende items, hetgeen in combinatie met het geringe aantal items een lage alpha-betrouwbaarheid in de hand werkt. De resultaten uit dit onderzoek zijn voor het overige vergelijkbaar met die van andere onderzoeken waarbij namelijk ook hoge alpha-betrouwbaarheden gevonden werden (Kulik & Kulik, 1974; Marsh, 1984). Uit de resultaten kan de conclusie getrokken worden dat groepen onderling samenhangende items, intern consistente schalen vormen. Deze schalen representeren de in de principale-componentenanalyse geïdentificeerde dimensies.

4.5 Studie 3: De betrouwbaarheid van studentoordelen.

4.5.1 Inleiding.

Het gebruik van beoordelingsvragenlijsten berust op de aanname dat studenten in staat zijn om op een nauwkeurige manier kenmerken van een object te observeren en om aan dat object numerieke waarden toe kennen. Een punt van zorg bij de gebruikers van evaluatieresultaten is, of studenten wel voldoende overeenstemming tonen bij de beoordeling van het onderwijs. Immers, studentoordelen worden door docenten pas serieus genomen als tussen studenten voldoende overeenstemming bestaat over het beoordeelde object.
In de nu volgende studie zal daarom de betrouwbaarheid van studentoordelen onderzocht worden. Centraal staat de vraag in hoeverre studenten homogene oordelen geven op de beoordelingsvragenlijst.
Studie 3 bestaat uit twee gedeelten. In het eerste deel wordt onderzoek beschreven dat tot doel had betrouwbaarheid van individuele oordelen te bepalen. In dat deel wordt de betrouwbaarheid van individuele studentoordelen binnen onderwijsgroepen berekend. Het tweede deel bevat de rapportage van onderzoek naar de betrouwbaarheid van geaggregeerde studentoordelen. In dit onderdeel staat de vraag centraal in hoeverre onderwijsgroepen bepaalde facetten van het onderwijs homogeen beoordelen.

4.5.2 Methode.

De betrouwbaarheid van studentoordelen is, zoals bleek in hoofdstuk 3, afhankelijk van de mate van overeenstemming tussen de studenten binnen een groep en van het aantal studenten in die groep. In studie 3 is sprake van twee onderzoeksgroepen waarin de betrouwbaarheid van oordelen over de kwaliteit van het onderwijs berekend wordt. De eerste onderzoeksgroep bestaat uit individuele studenten die in een willekeurig samengestelde onderwijsgroep (8-10 personen) het functioneren van de onderwijsgroep en van de tutor beoordelen. De tweede onderzoeksgroep bestaat uit onderwijsgroepen die oordelen geven met betrekking tot kwaliteitskenmerken van één blok.
In de eerste onderzoeksgroep geven individuele studenten een oordeel over kwaliteitskenmerken van probleemgestuurd onderwijs die specifiek zijn voor en afhankelijk van een enkele onderwijsgroep. De samenstelling van de onderwijsgroep en de door het faculteitsbureau toegewezen tutor zijn twee factoren die gedeeltelijk de kwaliteit van het onderwijs binnen een onderwijsgroep bepalen en die tevens onderwijsgroepen van elkaar laten verschillen. Binnen een onderwijsgroep komen studentoordelen niet onafhankelijk van elkaar tot stand. Studenten beïnvloeden elkaar wederzijds en worden op hun beurt weer beïnvloed door het gedrag van de tutor. Andersom geldt dat het gedrag van de tutor mede bepaald wordt door het

gedrag van de studenten in de onderwijsgroep. Vandaar dat individuele studentoordelen uitsluitend betrouwbaar zijn voor de beoordeling van die elementen van probleemgestuurd onderwijs die groepsgebonden zijn: het functioneren van de onderwijsgroep waarvan de betreffende studenten lid zijn, en het functioneren van de tutor.
In de tweede onderzoeksgroep vormen oordelen van onderwijsgroepen de eenheid van analyse. Per onderwijsgroep wordt het gemiddelde oordeel van alle studenten van de betreffende groep op een item of groep items berekend. Studentoordelen worden feitelijk geaggregeerd tot groepsoordelen. Deze oordelen zijn betrouwbaar voor het meten van die elementen van probleemgestuurd onderwijs die niet groepsafhankelijk zijn (bijvoorbeeld de kwaliteit van de taken, het skillslabprogramma, etc..).
Een veelgebruikte maat voor de betrouwbaarheid van oordelen is de "intraclass correlation coëfficiënt" (Feldman, 1977; Shrout & Fleiss 1979). Deze coëfficient is een betrouwbaarheidsindex die de ratio weergeeft van de ware variantie enerzijds en de som van de ware variantie en de errorvariantie anderzijds. Deze index relateert de variantie van oordelen binnen een groep aan de totale variantie van alle oordelen tussen groepen. Hoe hoger de intraclass correlation, des te kleiner de variantie binnen groepen ten opzichte van variantie tussen oordelen over alle groepen.
Om de diverse variantiecomponenten voor de intraclass correlatie te schatten, zowel voor het berekenen van de betrouwbaarheid van individuele oordelen als voor de betrouwbaarheid van de gemiddelde oordelen, is in deze studie een variantie-analyse model gebruikt dat in de literatuur omschreven wordt als 'split-plot design' (Kane & Brennan, 1977). In een dergelijk design zijn personen genest in een aantal groepen en over de groepen heen gekruist met items. Voor de berekening van de betrouwbaarheid van individuele studentoordelen volgens dit split-plot design, werden derhalve individuele studenten genest binnen onderwijsgroepen en over de onderwijsgroepen heen gekruist met items. De intraclass correlatie voor groepsoordelen werd berekend door onderwijsgroepen te nesten binnen blokken en over blokken heen te kruisen met items. Het structureel model is voor beide variantie-analytische designs hetzelfde. Individuele studentoordelen en groepsoordelen worden als random factor beschouwd, items of schalen als fixed. Onderstaande vergelijking geeft het structureel model weer.

$$x_{ij} = u + b_j + w_{ij}$$

x_{ij} = de geobserveerde score
u = het grote gemiddelde (grand mean)
b_j = itemeffect
w_{ij} = residuele component

In dit design zijn effecten van personen, interacties tussen personen en items, en random error niet onderscheidbaar. De residuele component w_{ij} bestaat uit de som van de persoonseffecten, interactie-effecten van persoon x item en de error-effecten. De voor de intraclass correlation benodigde variantie-componenten kunnen met behulp van een éénweg-variantie analyse (ANOVA) berekend worden. De intraclass coëfficiënt voor de betrouwbaarheid van één studentoordeel heeft in dit geval de volgende vorm (Shrout & Fleiss, 1979):

$$ICC_i = \frac{MSB - MSW}{MSB + (k-1)MSW}$$

ICC_i = intraclass correlation voor individuele oordelen
MSB = variantie tussen onderwijsgroepen
MSW = variantie binnen onderwijsgroepen, error-variantie
k = gemiddeld aantal beoordelaars binnen een onderwijsgroep

De ICC_i heeft een bovengrens van 1.0. Bij die waarde is er geen sprake van binnengroeps- of "within"variantie. De oordelen kunnen dan als perfecte replicaties van elkaar gezien worden. Naarmate de ICC meer nadert tot nul, duidt dit op toenemende niet-systematische verschillen tussen beoordelaars binnen dezelfde groepen (een toenemende binnengroepen-variantie), en dus op afnemende meetprecisie of betrouwbaarheid. Als de ICC_i gelijk is aan nul, betekent dat dat de variantie tussen groepen gelijk is aan de variantie binnen groepen. Het tot stand komen van oordelen berust dan louter op toeval. De ondergrens van de ICC_i is gedefinieerd als -1/k-1 en komt tot stand als MSB gelijk aan nul is.

4.5.3 Berekening betrouwbaarheid van individuele studentoordelen.

De betrouwbaarheid van een oordeel van een enkele student werd berekend door eerst de betrouwbaarheid van alle individuele studentoordelen binnen een onderwijsgroep te berekenen, hetgeen feitelijk overeenkomt met het berekenen van de betrouwbaarheid van het gemiddelde oordeel van studenten binnen een onderwijsgroep. Daarna werd, met behulp de Spearman-Brown formule voor testverkorting, de betrouwbaarheid van één studentoordeel berekend. Met behulp van de formule van Ebel (1951) werd de betrouwbaarheid van k individuele studentoordelen berekend, waarbij k het gemiddeld aantal personen binnen een groep weergeeft.

Ebel (1951) hanteert daarvoor een formule die als bijzonder geval van de hierboven genoemde formule gezien kan worden:

$$ICC_k = \frac{MSB - MSW}{MSB} = \frac{F - 1}{F}$$

k = gemiddeld aantal personen binnen een groep
MSB = variantie tussen groepen
MSW = variantie binnen groepen
F = MSB/MSW

De onderwijsgroepen in deze studie bevatten gemiddeld de oordelen van 7 studenten (k = 7). Per academiejaar werden éénweg variantie-analyses (ANOVA) uitgevoerd om de variantie van de items die betrekking hadden op de onderwijsgroep en de tutor, te splitsen in variantie tussen de groepen (de systematische variantie tussen onderwijsgroepen) en variantie binnen de groep(en) (de errorvariantie waarvan verondersteld wordt dat deze uit randomvariantie bestaat).

4.5.4 Berekening betrouwbaarheid groepsoordelen.

Eerst werden met behulp van de formule van Ebel (1951) de betrouwbaarheid berekend voor alle groepsoordelen binnen een blok (in feite het gemiddelde van alle groepsoordelen binnen een blok). Daarna werd door de Spearman-Brown formule toe te passen, de betrouwbaarheid van het oordeel van één groep groepsoordeel berekend. Per academiejaar werden éénweg-variantie-analyse (ANOVA) uitgevoerd om de benodige variantiecomponenten te berekenen. Variantie tussen blokken werd opgevat als systematische variantie. Variantie binnen blokken (de variantie van groepsoordelen binnen een blok) werd als errorvariantie beschouwd. In het academiejaar 1981/1982 bestond één blok uit gemiddeld 11 onderwijsgroepen. Het gemiddelde oordeel van alle onderwijsgroepen binnen één blok was dat jaar dus gebaseerd op 11 groepsoordelen. In de daaropvolgende academiejaren nam het aantal onderwijsgroepen toe tot een gemiddeld aantal van 18 onderwijsgroepen per blok (zie tabel 4.7). Om het aantal analyses in deze studie te beperken, werden als afhankelijke variabele uitsluitend de schaalscores van de dimensies van de vragenlijst gebruikt.

4.5.5 Resultaten: Deel I.

In tabel 4.6 zijn de resultaten uit de eerste analyse weergegeven. Per academiejaar worden de intraclass correlaties, gebaseerd op 7 individuele studenten en op 1 individuele student, weergegeven. De intraclass correlaties zijn berekend voor items die betrekking hebben op het functioneren van de onderwijsgroep en op het functioneren van de tutor (itemnummers v24 - v44: zie bijlage 1).

Tabel 4.6 Intraclass correlaties per item per academiejaar van individuele studentoordelen (N=1) en gemiddelde individue studentoordelen (N=7). De items hebben betrekking op het functioneren van de tutor en de onderwijsgroep.

Item	Academiejaar 1981/1982 N=7	N=1	1982/1983 N=7	N=1	1983/1984 N=7	N=1	1984/1 N=7	N
Onderwijsgroep.								
vr24 Ondwg.sys.proc.	.69	.24	.74	.29	.65	.21	.51	
vr25 Ondwg.zelfstudie	.70	.25	.75	.30	.66	.22	.68	
vr26 Ondwg.afspraken	--		.69	.24	.63	.20	.60	
vr27 Nakomen afspr.	.69	.24	.75	.30	.71	.26	.72	
vr28 Ondwg.prettig	--		.79	.35	.77	.32	.75	
vr29 Ondwg.productief	.74	.29	.79	.35	.77	.32	.77	
vr30 Actieve bijdrage	.69	.24	.75	.30	.75	.30	.74	
vr31 Ondwg.rem	--		.45	.10	.36	.07	.46	
vr32 Onafh.ondwg	--		--		--		.41	
Tutor.								
vr33 Ttr.begrip doelst.	.70	.25	.75	.30	.75	.30	.66	
vr34 Ttr.kennis ondsys.	.62	.19	.68	.23	.69	.24	.59	
vr35 Ttr.rol plezierig	.79	.35	.76	.31	.80	.36	.76	
vr36 Ttr.stim.werken	.69	.24	.64	.20	.70	.25	.61	
vr37 Ttr.disc.vragen	.78	.34	.72	.27	.72	.27	.69	
vr38 Ttr.disc.sturing	.85	.45	.80	.36	.82	.39	.77	
vr39 Ttr.stim.afspr.	.62	.19	.56	.15	.68	.23	.55	
vr40 Ttr.contrl.afspr.	.51	.13	.57	.16	.60	.18	.42	
vr41 Ttr.stim.inhds.	.72	.27	.75	.30	.77	.32	.69	
vr42 Ttr.stim.leerm.	--		.56	.15	.59	.17	.47	
vr43 Ttr.evaluatie.	.84	.43	.79	.35	.79	.35	.69	
vr44 Ttr.funkt.goed	.76	.31	.83	.41	.77	.32	.71	

Uit de resultaten in tabel 4.6 blijkt dat de betrouwbaarheic van de gemiddelde studentoordelen (n=7) redelijk hoog is. De intraclass correlaties hebben een waarde die rondom .70 ligt. De betrouwbaarheid van de gemiddelde oordelen op 4 van de 21 items, namelijk die op de items v31, v32, en v40 en v41, ligt in tegenstelling tot de overige 17 items, beneden .60. Met uitzondering van deze items blijkt dat het oordeel per studenten over het functioneren van de onderwijsgroep en de tutor over het geheel genomen redelijk betrouwbaar oordeel is. De betrouwbaarheid van één individueel studentoordeel bedraagt gemiddeld genomen ongeveer .25. Dat is tamelijk laag. Centra (1973), Feldman (1977), Marsh en Overall (1979) en Marsh (1982) melden vergelijkbare resultaten. Zij vonden eveneens dat de betrouwbaarheid van individuele oordelen varieert tussen .25 - .30. Deze onderzoekers rapporteren dat

betrouwbaarheid van gemiddelde studentoordelen (gebaseerd op 10 studenten) variëert tussen .70 - .75.Volgens deze auteurs is een betrouwbaarheid van plusminus .70 acceptabel.
In tabel 4.6 worden per academiejaar intraclass correlaties gerapporteerd. Uit de resultaten blijkt dat het merendeel van de intraclass correlaties tamelijk stabiel is over de afzonderlijke academiejaren heen. Bij sommige items (bijvoorbeeld item 24) is echter sprake van tamelijke sterke fluctuaties. Dergelijke fluctuaties kunnen een artefact zijn van de wijze waarop de intraclass correlatie-coëfficiënt berekend wordt. Als binnen een academiejaar minder spreiding tussen de groepen is dan in andere academiejaren zal, bij een gelijkblijvende spreiding binnen groepen, de intraclass correlatie lager worden.

4.5.6 Resultaten: Deel 2.

In tabel 4.7 worden de resultaten van de tweede analyse weergegeven. Per academiejaar worden per schaal de intraclass correlaties gerapporteerd voor singuliere groepsoordelen en voor successievelijk 11, 13, 15, en 18 groepsoordelen.

Tabel 4.7 Intraclass correlaties per schaal per academiejaar van singuliere groepsoordelen (N=1) en gemiddelde groepsoordelen (N=11, N=13, N=15, N=18).

Dimensie	Academiejaar 1981/1982 N=11	N=1	1982/1983 N=13	N=1	1983/1984 N=15	N=1	1984/1985 N=18	N=1
1 Taken	.95	.63	.91	.44	.94	.51	.90	.33
2 Tutor	-.67	-.04	.39	.05	.43	.05	.18	.03
3 Onderwijsgroep	.78	.24	.60	.10	.51	.06	.69	.11
4 Skillslab	--	--	--	--	.72	.15	.94	.47
5 Zwaarte	.88	.40	.95	.59	.95	.56	.98	.73
6 Bloktoets	.82	.29	.84	.29	.81	.22	.85	.24
7 Leermiddelen	.85	.34	.89	.38	.94	.51	.94	.47
8 Globaal oordeel	.86	.36	.90	.41	.88	.33	.95	.51
10 Raadplegen inh.desk.	.18	.02	.57	.09	.51	.06	.32	.03
11 Soc. Vaardigheden	--	--	--	--	.73	.15	.89	.31
12 Onafhankelijk stud.	--	--	.36	.04	.61	.09	.61	.08
13 Structuring Onderw.	.50	.08	.53	.08	.71	.14	.67	.10
14 Afwisseling	.84	.32	.95	.59	.89	.59	.87	.27
15 Tutor harg werken	-.24	-.02	.46	.06	.42	.05	-.28	-.01

Uit de resultaten in tabel 4.7 blijkt dat de betrouwbaarheid van gemiddelde groepsoordelen zeer hoog is op de schalen met betrekking tot de kwaliteit van de taken, het skillslab, leerstofkenmerken, en de bloktoets, namelijk gemiddeld .90. Dit zijn schalen die betrekking hebben op elementen van het onderwijsprogramma die voor alle onderwijsgroepen gelijk

zijn. Verschillende onderwijsgroepen beoordelen in dit geval dus dezelfde objecten of programma-elementen. De betrouwbaarheid van één groepsoordeel is op deze schalen groot genoeg om als redelijk bruikbare indicator te dienen voor bijvoorbeeld het meten van de kwaliteit van de taken. Deze resultaten suggereren dat onderwijsgroepen een homogeen oordeel geven over programma-elementen die voor alle studenten in een blok hetzelfde zijn.
Voorts blijkt uit de resultaten in tabel 4.7 dat zowel de betrouwbaarheid van gemiddelde groepsoordelen als van één groepsoordeel vrijwel nihil is als deze oordelen betrekking hebben op programma-elementen die groepsspecifiek zijn. Dit geldt met name voor aspecten van tutorgedrag in deze. Deze resultaten zijn verklaarbaar uit het feit dat verschillende onderwijsgroepen altijd een andere tutor hebben. Verschillende groepen beoordelen theoretisch verschillende objecten. In dergelijke gevallen is het zelfs een vereiste dat de betrouwbaarheid van deze oordelen laag is. Uit het feit dat de betrouwbaarheid van deze oordelen aan deze vereiste voldoet, kan men concluderen dat onderwijsgroepen blijkbaar verschillende tutoren anders beoordelen. Deze resultaten suggereren bovendien dat aan een andere conditie ten aanzien van de gewenste psychometrische kwaliteiten van oordelen wordt voldaan: oordelen moeten kunnen discrimineren tussen objecten die voor beoordelaars verschillend zijn en tevens moeten ze kunnen convergeren voor objecten die voor beoordelaars identiek zijn. Deze conditie kan men zien als een validiteitseis voor de kwaliteit van oordelen. In studie 5, waarin de constructvaliditeit van studentoordelen onderzocht is, wordt nader aandacht aan deze conditie besteed. Tenslotte blijkt uit de resultaten in tabel 4.7, dat de betrouwbaarheid van de gemiddelde groepsoordelen redelijk hoog is op de onderwijsgroepschaal. De betrouwbaarheid van één groepsoordeel is daarentegen laag. De enige interpretatie die redelijk lijkt, is dat er gemiddeld genomen geen grote verschillen zijn in de wijze waarop verschillende onderwijsgroepen zichzelf binnen een blok beoordelen, maar dat deze verschillen wel zodanig zijn dat het oordeel van één onderwijsgroep geen bruikbare indicator is voor het functioneren van alle onderwijsgroepen in een blok.

4.5.7 Deel 1 en deel 2: discussie en conclusies.

Psychometrische theorieën over het meten van kenmerken van objecten met behulp van beoordelaars veronderstellen dat iedere beoordelaar in principe dezelfde ware score aan een object toekent (Guilford, 1954; Larson, 1979). Als de beoordelaars onzorgvuldig hun werk doen (random meetfouten) of beïnvloed worden door niet aan dat object gerelateerde factoren (systematische meetfouten), zal de geobserveerde score niet langer bestaan uit louter ware score maar tevens componenten bevatten die afkomstig zijn van genoemde fouten.
In deel 1 van deze studie werd gevonden dat één studentoor-

deel een lage betrouwbaarheid heeft. Vanuit dat oogpunt bezien is het oordeel van één willekeurige student over het functioneren van de onderwijsgroep en de tutor, slechts van zeer beperkte waarde. Daar staat tegenover dat het gemiddelde oordeel van alle individuele studenten binnen een onderwijsgroep een tamelijk hoge betrouwbaarheid heeft. In deel 2 van deze studie kwam naar voren, dat één groepsoordeel al een redelijke hoge betrouwbaarheid bezit als het gaat om vragenlijstschalen die betrekking hebben op elementen van het onderwijs die voor alle onderwijsgroepen hetzelfde zijn (blokboek, bloktoets, skillslabprogramma, globaal oordeel, en moeilijkheidsgraad). De betrouwbaarheid van de gemiddelde groepsoordelen is op deze schalen zelfs zeer hoog.
Deze resultaten lijken twee conclusies te rechtvaardigen. In de eerste plaats blijkt dat het gemiddelde oordeel van alle studenten binnen een onderwijsgroep, een betrouwbare indicator vormt voor het functioneren van de onderwijsgroep en de tutor. In de tweede plaats kan geconcludeerd worden dat gemiddelde groepsoordelen van alle onderwijsgroepen binnen een blok, een zeer bruikbare maat zijn voor de beoordeling van de kwaliteit van programma-elementen die alle onderwijsgroepen betreffen.
Bij deze conclusies moet de volgende kanttekening geplaatst worden. Theoretisch gezien kunnen de in deze studie gevonden hoge betrouwbaarheden en de daaraan verbonden conclusies enigszins afgezwakt worden. Feldman (1977) constateert in dat verband dat studentoordelen meestal gebaseerd zijn op retrospectieve observaties. Bij de beoordeling van het onderwijs wordt met andere woorden een sterk beroep gedaan op het geheugen van studenten. Op het einde van een onderwijseenheid geven studenten immers een oordeel dat gebaseerd is op een reeks observaties die eerder in de tijd plaatsvonden. Daarom bestaat volgens Feldman (1977) steeds de mogelijkheid, dat de op deze wijze verkregen oordelen niet compleet onafhankelijk zijn van eerdere ervaringen in het onderwijs en van medestudenten. Als deze redenering toegepast wordt op oordelen van studenten binnen onderwijsgroepen, bestaat het risico dat de variantie binnen groepen kleiner is dan men in werkelijkheid zou verwachten. Dit zou tot gevolg kunnen hebben dat, onder de assumptie dat binnengroepsvariantie uitsluitend uit random meetfouten bestaat, een overschatting van de betrouwbaarheid ontstaat. Noch Feldman noch andere onderzoekers hebben empirische indicaties in deze richting gevonden (Marsh & Cooper, 1981). In dat licht lijkt de conclusie des te meer gerechtvaardigd dat de resultaten uit deze studie op een hoge betrouwbaarheid van studentoordelen wijzen.

4.6 Studie 4: Criteriumvaliditeit van studentoordelen.

4.6.1 Inleiding.

In de nu volgende studie wordt onderzoek beschreven waarin studentoordelen gevalideerd worden aan het studietoetscriterium. De gehanteerde onderzoeksopzet voldoet, zoals dadelijk zal blijken, aan de eisen door Cohen (1981) gesteld aan goed validatie-onderzoek. Bloktoetsen worden in deze validiteitsstudie gebruikt als extern criterium.

4.6.2 Methode.

De onderzoeksgroep bestond uit 1063 onderwijsgroepen die verdeeld waren over 4 academiejaren (1981/1982 tot en met 1984/1985). Ieder academiejaar bestond uit vier studiejaren met in totaal 20 blokken. Uiteindelijk zijn 76 blokken in de analyse betrokken. Van vier blokken waren geen gegevens beschikbaar. Eenheid van analyse was de onderwijsgroep. Per onderwijsgroep werden evenals in de vorige studies gemiddelde oordelen (gemiddelde oordeel per onderwijsgroep) berekend over de in deze studies beschreven schalen. Studenten waren random ingedeeld in onderwijsgroepen.
Per bloktoets werden goed-scores getransformeerd in percentuele scores om vergelijkingen tussen toetsen mogelijk te maken. Voor elke onderwijsgroep werden gemiddelde percentuele toetsscores berekend. Verdere analyses werden steeds op deze groepsgemiddelden uitgevoerd. Tutoren droegen geen verantwoordelijkheid voor de constructie van de bloktoets.
Deelname aan de toetsen was verplicht. De toetsresultaten hadden in het academiejaar 1981/1982 geen consequenties voor de studenten in termen van zakken of slagen. In de academiejaren 1982/1983 en 1983/1984 hadden de toetsresultaten wel consequenties: voortgangstoetsresultaten met de score 'twijfelachtig' konden gecompenseerd worden met een voldoende resultaat op de bloktoets. In het academiejaar 1984/1985 hadden bloktoetsresultaten dezelfde consequenties als in beide voorgaande academiejaren, doch uitsluitend voor het eerste studiejaar (propedeutisch examen) en het vierde studiejaar (doctoraal examen). Per academiejaar werden voor de afzonderlijke studiejaren produkt-moment correlaties berekend tussen de groepsoordelen en bloktoetsscores.

4.6.3 Resultaten.

In tabel 4.10 zijn de produkt-moment correlaties tussen groepsoordelen en bloktoetsscores weergegeven. Uit de resultaten blijkt dat in het academiejaar 1981/1982 groepsoordelen op verschillende schalen positief correleren met bloktoetsscores. De takenschaal correleert bijvoorbeeld in het eerste studiejaar van dit academiejaar, .64 met de bloktoetsscores. Binnen dit academiejaar blijkt geen consistent patroon waarneembaar in de correlaties tussen groepsoordelen

en bloktoetsscores. De takenschaal correleert bijvoorbeeld alleen in het eerste studiejaar positief met de bloktoetsscores. Uit de resultaten in de daaropvolgende academiejaren blijkt dat evenmin sprake is van consistente correlatiepatronen.
In het academiejaar 1982/1983 en het daaropvolgende academiejaar 1983/1984 worden zelfs beduidend minder significante correlaties gevonden dan in het academiejaar 1981/1982. In het academiejaar 1984/1985 is sprake van een toegenomen aantal correlaties. Met name in het tweede studiejaar blijkt dat het merendeel van de schalen positief correleert met de bloktoetsscores. De vraag die zich uiteraard voordoet is hoe deze resultaten geïnterpreteerd kunnen worden.

4.6.4 Discussie en conclusies.

In hoofdstuk 3 werden verschillende hypothesen besproken voor het verklaren van correlaties tussen de studentoordelen en toetsprestaties. Drie hypothesen bleken het meest voor de hand te liggen. De eerste hypothese was dat correlaties tussen studentoordelen en studieprestaties een bewijs zijn voor de validiteit van studentoordelen, omdat studenten meer leren naarmate de kwaliteit van het onderwijs beter is. Volgens de tweede hypothese kunnen deze correlaties echter verklaard worden door het zogenaamde 'grading - leniency effect': studenten geven positieve oordelen over het onderwijs als zij van docenten voor een tentamen een hoog cijfer verwachten of ontvangen. Naarmate een docent soepeler is in het toekennen van hoge cijfers zullen de studentoordelen dus hoger worden. Dergelijke correlaties zijn volgens deze hypothese juist te wijten aan een gebrek aan validiteit van studentoordelen.
De derde hypothese verklaart substantiële correlaties door te wijzen op de effecten van 'ability-rating bias'. Indien groepen voorafgaande aan de cursus reeds verschillen wat betreft hun voorkennis of motivatie, dan kunnen aan het einde van de cursus verschillen in studieprestaties verklaard worden door verschillen vooraf die te wijten zijn aan de groepssamenstelling.
De resultaten uit deze studie kunnen niet verklaard worden met behulp van de hypothese betreffende de "ability-rating bias" (zie paragraaf 3.7.2). Immers, studenten worden aselect in onderwijsgroepen ingedeeld, waardoor men verwachten mag dat (met uitzondering van toevalstreffers) er geen wezenlijke verschillen tussen onderwijsgroepen bestaan aan het begin van een blok.

Tabel 4.10: Correlaties tussen groepsoordelen en procentuele goed-score de bloktoets.

		Academiejaar															
		1981				1982				1983				1984			
Studiejaar		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	Dimensie																
1	Taken.	64"	11	28	21	04	-13	-10	-09	10	-13	-06	-30	-06	30"	-05	-13
2	Tutor	03	31	01	-09	07	06	19	-02	-09	-04	-22	11	16	14	-11	13
3	Ondwg	46"	25	41	48'	-07	-06	-15	19	01	09	-08	-05	08	32"	21	32'
4	Skillslab	--	--	--	--	--	--	--	--	-19	03	-04	-13	-14	31"	19	13
5	Zwaarte	23	06	53'	25	-07	-20	-08	41'	-03	01	-22	23	-01	40"	03	15
6	Bloktoets	56"	40'	03	44	28'	05	03	38'	03	10	05	17	20	48"	23	52"
7	Leermidd.	33'	27	-34	39	00	-34'	20	-12	-16	-11	03	05	-24'	55"	08	-12
8	Glb.ordl.	51"	37'	57"	21	-19	-31'	-20	34	20	02	-21	15	-01	44"	04	32'
10	Inhoudsd.	00	08	00	17	-03	11	06	-07	-06	-14	-16	-02	-01	44"	04	32'
11	Soc.vaard.	--	--	--	--	--	--	--	--	-30'	08	04	18	00	37"	15	18
12	Onafhank.	--	--	--	--	11	24	-03	-21	-13	15	14	-14	-11	07	-08	-12
13	Struct.	-11	08	15	09	27'	07	-18	08	-10	-07	10	05	05	20	-22	16
14	Afwiss.	33'	-07	39	-12	-42"	-11	-30	36	15	-07	-14	16	-11	23'	07	33'
15	Tutor hw	07	-06	08	01	00	06	22	-27	-05	00	-06	-05	08	01	01	-01
	N	72	45	30	24	91	59	47	39	87	89	59	59	103	108	72	73

Noot: correlaties zijn vermenigvuldigd met 100.
Significantiegrenzen: ' p<.01
" p<.001

Volgens de hypothese over het zogenaamde "grading-leniency effect", die correlaties tussen (verwachte) tentamenresultaten en studentoordelen verklaart in situaties waarin tentamenresultaten bepalend zijn voor de studievoortgang van studenten, zou men in deze studie een groot aantal correlaties mogen verwachten tussen het merendeel van de schalen van de vragenlijst en de scores op de bloktoets. Studenten vullen de beoordelingsvragenlijst immers na afloop van de bloktoets in en worden volgens deze hypothese beïnvloed door de verwachte of behaalde prestaties. Uit tabel 4.10 blijkt echter dat voornamelijk in het academiejaar 1981/1982 en grotendeels in het tweede studiejaar van het academiejaar 1984/1985 positieve correlaties gevonden worden. Prestaties op de bloktoets hadden voor beide studentgroepen geen consequenties voor de studievoortgang van studenten. Volgens de genoemde hypothese zou men echter tegengestelde resultaten mogen verwachten: significante positieve correlaties in studentjaargangen waarvoor de bloktoets wel consequenties heeft (namelijk compensatie, als de uitslag op de voortgangstoets 'twijfelachtig' is) en geen significante correlaties in jaargangen waarvoor de compensatieregel niet geldt. De onderzoeksresultaten geven dus geen steun aan de veronderstelling van meetfouten ten gevolge van een grading-leniency effect.
Resteert de hypothese dat studentoordelen over het onderwijs valide zijn bij substantiële correlaties tussen bepaalde schalen van de vragenlijst en de bloktoets. Althans wat betreft die schalen waarvan het theoretisch aannemelijk is dat zij correleren met studentleren. Dit lijkt het geval te zijn in het academiejaar 1981/1982. Op een aantal schalen (taken, blokboek, onderwijsgroep, aansluiting van de bloktoets, leermiddelen, globaal oordeel, en variatie) worden redelijk positieve correlaties gevonden die tevens plausibel zijn. Deze correlaties hebben vergelijkbare waarden (rond .45) als in de meta-analyse van Cohen (1981). De schaal 'globaal oordeel' correleert ongeveer .45 met de bloktoets. De beoordeling van het functioneren van de onderwijsgroep correleert in twee studiejaren ongeveer .40 met de bloktoetsresultaten. Bovendien worden op de schalen die betrekking hebben op het functioneren van de tutor, geen correlaties gevonden tussen studentoordelen en toetsprestaties. Theoretisch gezien worden bij laatstgenoemde schalen geen significante correlaties verwacht omdat de tutor niet de klassieke docentenrol vervult in de betekenis van het direct overbrengen van kennis (zie hoofdstuk 2), maar de opdracht heeft om als begeleider van de groep op te treden. De resultaten uit het academiejaar 1981/1982 suggereren dan ook dat studentoordelen in dat jaar een redelijk goede criteriumvaliditeit bezitten. Uit de resultaten in tabel 4.10 blijkt echter dat in de daaropvolgende academiejaren de samenhang tussen studentoordelen en toetsprestaties grotendeels verdwenen is. De meeste correlaties in de volgende academiejaren zijn niet significant. Voor dit fenomeen lijken drie verklaringen voor de hand te liggen.

De eerste verklaring is dat de bloktoets in de daaropvolgende jaren geen goede weerspiegeling vormde van het leerproces van studenten. Deze verklaring kan makkelijk van de hand gewezen worden. De wijze waarop de bloktoets en het leerproces in deze periode tot stand kwam, verschilde niet wezenlijk van voorgaande jaren. De tweede verklaring is dat in de betreffende periode de vragenlijst geen valide indicator vormde voor het meten van de kwaliteit van het onderwijs. Ook deze verklaring is weinig doeltreffend. Immers uit validatie-onderzoek bleek dat de vragenlijst ook in deze jaren een valide indicator was (zie paragraaf 4.6 en 4.7). Resteert als verklaring dat de relaties tussen bloktoetsprestaties en dimensies van probleemgestuurd onderwijs veranderd zijn. Twee aannamen worden daarbij gehanteerd. De eerste betreft dat de psychometrische kwaliteit van beide instrumenten binnen de onderzoeksperiode ongewijzigd blijft. De tweede aanname is dat bloktoetspresaties mede afhankelijk zijn van de kwaliteit van het gegeven onderwijs.
Het verdwijnen van de samenhang tussen studentoordelen over de kwaliteit van probleemgestuurd onderwijs en bloktoetsresultaten kan met grote waarschijnlijkheid duiden op een verandering van studiestrategie van studenten. Veranderingen in de relatie bloktoetsprestaties en studentoordelen op de vragenlijst kunnen dan een indicatie vormen voor veranderingen in studiegedrag van studenten. Leerdoelen van onderwijsgroepen zouden bijvoorbeeld een meer of minder belangrijke rol kunnen spelen bij de studie-activiteiten van studenten. De onderwijsgroep en de resultaten van het werk dat in die groep gedaan wordt, spelen in een dergelijke interpretatie geen rol (meer) in de manier waarop studenten hun studie aanpakken. De resultaten op de bloktoets zouden dan niet meer verklaard worden uit een studiestrategie waarin het met anderen samenwerken aan taken een plaats heeft, maar uit een strategie waarover de vragenlijst geen informatie verschaft. Deze verklaring lijkt des te meer aannemelijk als men bedenkt dat de bloktoets vanaf het academiejaar 1982/1983 een andere status kreeg. De bloktoets kon vanaf dat moment de score 'twijfelachtig' op de voortgangstoets compenseren. In het voorgaande werd geconstateerd dat het merendeel van de correlaties tussen studentoordelen en toetsprestaties gevonden werd bij studentgroepen waarvoor de bloktoets geen consequenties had. Als studenten uit de latere jaargroepen hun studiestrategie wijzigden om een voldoende resultaat op de bloktoets te behalen, kan dat een verklaring vormen voor de niet-significante correlaties tussen studentoordelen en toetsprestaties. Immers in de vragenlijst wordt van de veronderstelling uitgegaan dat leerdoelen die door de onderwijsgroep geformuleerd worden, het uitgangspunt vormen voor de zelfstudie van studenten. Bij een gewijzigde studiestrategie hoeft dat niet langer of althans minder het geval te zijn. De vragenlijst meet volgens deze verklaring wel het functioneren van probleemgestuurd onderwijs, maar bevat geen indicatoren die toetsprestaties voorspellen. Onderzoek naar deze proble-

matiek (Gijselaers, Schmidt & Wijnen, 1984) heeft overigens geen bevredigend antwoord opgeleverd.
Samenvattend kan geconcludeerd worden dat het onderzoek naar de samenhang tussen studentoordelen en toetsprestaties geen verduidelijking biedt in de criteriumvaliditeit van de vragenlijst.

4.7 Studie 5: Onderzoek naar de constructvaliditeit van studentoordelen: een multitrait-multimethod analyse.

4.7.1 Inleiding.

Uit de vorige studie bleek dat de criteriumvalidatie benadering gemengde resultaten oplevert, gezien de tamelijk lage correlaties tussen het criterium en de studentoordelen. In deze studie zal daarom een constructvalidatie-benadering toegepast worden teneinde meer inzicht te krijgen in de validiteit van studentoordelen. Een multitrait-multimethod analyse wordt verricht om de konvergerende en divergerende validiteit van de beoordelingsvragenlijst voor studenten vast te stellen. Naast studentoordelen werden in deze studie tutoroordelen verzameld over de kwaliteit van het onderwijs. Het inwinnen van student- en tutoroordelen kunnen als twee verschillende dataverzamelingsmethoden beschouwd worden die ieder een aantal gemeenschappelijke kwaliteitskenmerken van probleemgestuurd onderwijs meten.
Voorafgaande aan de multitrait-multimethod analyse worden met behulp van principale componentenanalyse, de onderliggende dimensies van de vragenlijsten voor tutoren en studenten onderzocht. Congruente dimensies vormen namelijk een eerste indicatie voor de constructvaliditeit van studentoordelen. Als maatstaf voor de constructvaliditeit worden vervolgens de konvergerende en divergerende validiteit, van studentoordelen bepaald.

4.7.2 Onderzoeksgroep.

In het academiejaar 1984/1985 werden zowel studentoordelen als tutoroordelen verzameld. Aan het einde van ieder blok werden aan studenten en tutoren vragenlijsten voorgelegd die voor een deel, 32 gemeenschappelijke items bevatten. In voorgaande academiejaren bevatten beide vragenlijsten te weinig gemeenschappelijke items (20) om zinvol multitrait-multimethod onderzoek mee te verrichten. Vandaar dat de onderzoeksgroep in deze studie beperkt is tot het academiejaar 1984/1985 waarbij van 357 onderwijsgroepen de studentoordelen beschikbaar waren. Van 176 onderwijsgroepen zijn zowel de studentoordelen als de tutoroordelen beschikbaar. In deze studie zullen alleen de gegevens gebruikt worden van onderwijsgroepen waarvan zowel studentoordelen als tutoroordelen beschikbaar zijn. Studentoordelen worden in deze studie geaggregeerd tot groepsoordelen.

4.7.3 Analyse.

Op beide vragenlijsten werd een principale-componentenanalyse uitgevoerd over de items (32) die voor beide lijsten gemeenschappelijk waren. Principale componenten werden geextraheerd op basis van het Kaiser-criterium (de eigenwaarde van de component moet groter of gelijk aan 1.0 zijn). Na de extractie werden de componenten oblique geroteerd. Op basis van de uitkomsten van deze analyse werden schalen geconstrueerd die voor beide vragenlijsten identiek waren. Daarna werden correlaties tussen deze schalen berekend teneinde een multitrait-multimethod analyse te kunnen verrichten. Tevens werden voor iedere schaal de alpha-betrouwbaarheden berekend.

4.7.4 Resultaten.

In tabel 4.11 zijn voor beide vragenlijsten de eigenwaarden van de componenten, de percentages verklaarde variantie en de cumulatieve percentages verklaarde variantie weergegeven. In de principale-componentenanalyse voor de studentvragenlijst werden 8 componenten gevonden die samen 71.8 procent van de totale variantie verklaren. De principale-componentenanalyse voor de tutorvragenlijst leverde 9 componenten op die samen 65.7 procent van de variantie verklaren.

Tabel 4.11 Studentenvragenlijst (32 items). 176 cases
Eigenwaarden en percentage verklaarde variantie van de componenten.

	Studentvragenlijst			Tutorvragenlijst		
Component	Eigen-waarde	Percentage variantie	Cumulatief percentage	Eigen-waarde	Percentage variantie	Cumulatief percentage
1	9.2	27.8	27.8	8.4	25.4	25.4
2	4.1	12.3	40.1	3.5	10.5	36.0
3	3.0	9.0	49.1	2.3	6.9	42.8
4	2.3	7.1	56.2	2.1	6.4	49.2
5	1.5	4.6	60.8	1.6	4.7	54.0
6	1.4	4.2	65.0	1.5	4.5	58.5
7	1.2	3.6	68.6	1.3	4.0	62.5
8	1.0	3.1	71.8	1.1	3.4	65.9
9				1.0	2.8	71.7

In tabel 4.12 is de factorpatroonmatrix van de student- en tutorvragenlijst weergegeven. Alleen ladingen groter of gelijk aan .30 en ladingen kleiner dan of gelijk aan -.30 zijn in de matrix opgenomen.
De factorpatroonmatrix van de studentvragenlijst bestaat uit 8 dimensies. De meeste items laden slechts op één dimensie. De gevonden dimensies zijn goed interpreteerbaar. Ze hebben achtereenvolgens betrekking op de studielast in het blok (F1), de rol van de tutor (F2), het functioneren van de onderwijsgroep (F3), de kwaliteit van de taken (F4), de vakinhoudelijke bijdrage van de tutor in de onderwijsgroep (F5), het onafhankelijk van de onderwijsgroep studeren (F6), het aantal taken (F7), en de variëteit in taakvormen en keuze van onderwerpen (F8).
De factorpatroonmatrix van de tutorvragenlijst bestaat uit 9 dimensies. Ook deze dimensies zijn redelijk goed interpreteerbaar. Ze hebben achtereenvolgens betrekking op de afwisseling in het aanbod van taken en onderwerpen (F1), de onderwijsgroep (F2), de studielast in het blok (F3), de vakinhoudelijke bijdrage (F4), het toezien van de tutor op de voortgang van de onderwijsgroep (F5), het onafhankelijk van de onderwijsgroep studeren (F6), de kennis en het inzicht van de tutor in probleemgestuurd leren (F7), de kwaliteit van de taken (F8), en de beschikbaarheid van leer- en hulpmiddelen bij de taken in het blokboek (F9).
Uit de gegevens in de matrix blijkt echter dat aan de tutorvragenlijst, ondanks het feit dat gewerkt wordt met dezelfde items als in de studentvragenlijst, een aantal inhoudelijk andere dimensies ten grondslag liggen. Studielast wordt door tutoren bijvoorbeeld gecombineerd met de mate van aansluiting van het blok op de voorkennis van studenten en met het aantal taken. Voor studenten is studielast afhankelijk van de leerstof waarnaar de taken verwijzen en vormt het aantal taken een aparte beoordelingsdimensie. Tutoren koppelen het functioneren van de onderwijsgroep bijvoorbeeld aan hun eigen functioneren en aan de mate waarin zij het maken van afspraken niet stimuleerden.
Uit de gegevens blijkt de tutordimensies slechts gedeeltelijk congruent zijn met de studentdimensies. Feitelijk is alleen de tweede beoordelingsdimensie (F2) uit de tutorvragenlijst identiek aan een dimensie uit de studentvragenlijst, namelijk de dimensie die betrekking heeft op het functioneren van de onderwijsgroep. De andere dimensies uit de tutorvragenlijst komen slechts gedeeltelijk overeen met de dimensies van de studentvragenlijst.

Tabel 4.12: Factorpatroonmatrix van student- en tutorvragenlijst.

Items	STUDENT								TUTOR							
	F1	F2	F3	F4	F5	F6	F7	F8	F1	F2	F3	F4	F5	F6	F7	F8
1 Studielast																
v21-t17	|86|										|-77|		30			
v04-t04	|85|										|-66|					
v05-t05	67															64
v20-t16	52						35									
v15-t11	51			50												35
2 Tutor																
v37-t31		86										73				
v39-t37		84								61			52			
v44-t43		83								-51		35				
v33-t30		|81|												|-86|	
v34-t29		|80|												|-96|	
v43-t38		62			59								83			
3 Onderwijsgroep																
v27-t23			|-84|							|-81|					
v30-t25			|-87|							|-79|					
v29-t24			|-86|							|-80|					
v28-t26			|-85|							|-78|					
v01-t01			|-71|	-31						|-75|					
v26-t22			|-64|							|-50|				-32	
v24-t21			|-45|		44					|-58|					
4 Taken																
v12-t08				|-89|					37							|57|
v11-t07				|-88|					34					30		|37|
v10-t06				-80						36						
v17-t13				-68						36						
v14-t10				|-68|												|64|
v22-t18				-54										47		
v03-t03				|-44|					31							|37|
5 Tutorbijdrage aan discussie																
v38-t34		32			|-82|							|82|				
6 Onafhankelijk blok.																
v16-t12						|79|							|86|		
v32-t28						|68|							|88|		
7 Aantal taken																
v23-t19							|80|					|-49|			
v02-t02							35				-43	32				
8 Variëteit taken																
v13-t09								-68								
v19-t15								|64	89|							
v18-t14								|56	80|							

Noot. De getallen in de tabel zijn met 100 vermenigvuldigd. Itemladinge in de hokjes zijn ladingen van items die opgenomen zijn binnen de schalen voor de MIMM-analyse.

4.7.5 Discussie en conclusies.

De principale-componentenanalyse die verricht werd op twee gemeenschappelijke itemsets, had enerzijds ten doel eventuele overeenkomsten in beoordelingsdimensies van studenten en tutoren te identificeren en had anderzijds ten doel om schalen te construeren die gebruikt konden worden voor de multitrait-multimethod analyse. Uit de resultaten blijkt dat slechts één gemeenschappelijke dimensie geïdentificeerd kan worden, namelijk de dimensie die betrekking heeft op het functioneren van de onderwijsgroep. De andere dimensies laten slechts een gedeeltelijke overlap zien. Deze resultaten zijn nauwelijks in overeenstemming met resultaten uit vergelijkbaar onderzoek van Marsh, Overall en Kesler (1979) en Marsh (1982). Deze onderzoekers gaven docenten eveneens een vragenlijst die identieke items bevatte als de vragenlijst voor studenten. Het enige verschil was dat de items in de ik-vorm geformuleerd waren. Uit hun onderzoek bleek dat docenten bijna identieke beoordelingsdimensies hanteren als studenten. Een verklaring voor de in studie 5 gevonden verschillen met de onderzoeksuitkomsten van Marsh c.s., kan gelegen zijn in het feit dat in de medische faculteit tutoren geen vakinhoudelijke experts hoeven te zijn en het feit dat tutoren geen docentenrol vervullen die vergelijkbaar is met de rol van docenten in de onderzoeken van Marsh. Dit kan ertoe leiden dat tutoren bijvoorbeeld zaken als de kwaliteit van de taken anders inschatten dan studenten. Het is echter opmerkelijk dat studenten en tutoren wel een congruente beoordelingsdimensie hanteren ten aanzien van het functioneren van onderwijsgroepen. Samenvattend kan geconcludeerd worden dat de beoordelingsdimensies die studenten en tutoren hanteren met betrekking tot probleemgestuurd onderwijs, slechts gedeeltelijk overeenstemmen.

4.7.6 Multitrait-multimethod analyse.

Op basis van de uitkomsten van de principale-componentenanalyse op de studentvragenlijst werden schalen geconstrueerd die identieke items voor beide vragenlijsten bevatten. Alleen items van de tutorvragenlijst die op vergelijkbare dimensies van beide lijsten voorkwamen, werden in de gemeenschappelijke schalen opgenomen. Voor beide vragenlijsten werden zeven schalen geconstrueerd. Items van dimensie 7 van de studentvragenlijst (aantal taken) en de daarbij corresponderende items van de tutorvragenlijst werden niet tot één schaal omgevormd. In tabel 4.13 is aangegeven welke items in de schalen werden opgenomen. Schaalscores werden berekend door de itemsscores te middelen. Vervolgens werden tussen de schalen van de beide vragenlijsten produkt-moment correlaties berekend.

4.7.7 Resultaten.

Campbell en Fiske (1959) bepleiten een constructvalidatiebenadering waarin met behulp van multitrait-multimethod analyse de validiteit van een meetinstrument onderzocht wordt. Met behulp van deze analysemethode kan de convergerende en divergerende validiteit van een meetinstrument bepaald worden waardoor uitspraken gedaan kunnen worden over de constructvaliditeit van het instrument. De constructvaliditeit van een meetinstrument is afhankelijk van de mate waarin hoge correlaties gevonden worden met instrumenten die geacht worden hetzelfde te meten (convergerende validiteit), en van de mate waarin nulcorrelaties of lage correlaties gevonden worden met instrumenten die traits meten die geen relatie hebben met het te valideren instrument (divergerende validiteit).
In deze studie werd de kwaliteit van het onderwijs gemeten met behulp van twee methoden: student- en tutoroordelen. De traits zijn in dit geval dimensies van probleemgestuurd onderwijs (tutor, blokboek en taken, onderwijsgroep, etc.). De dimensies zijn geoperationaliseerd door schalen te construeren die gebaseerd zijn op de uitkomsten van de in deze studie beschreven principale-componentenanalyse.
Berekening van de correlaties tussen traits en methoden levert een multitrait-multimethod correlatiematrix op. De correlatiematrix in tabel 4.13 omvat 14 rijen en 14 kolommen. De hoofddiagonaal bestaat uit twee betrouwbaarheidsdiagonalen. Op deze diagonalen staan, met uitzondering van schaal 5 (deze schaal bestaat uit 1 item), de alpha-betrouwbaarheden van de schalen vermeld. Aangrenzend aan de betrouwbaarheidsdiagonalen liggen de heterotrait-monomethod driehoeken. Hierin staan de correlaties tussen de traits voor beide methoden. De in de tabel onderstreepte correlaties staan op de zogenaamde validiteitsdiagonaal. Deze correlaties geven de convergerende valditeit van de instrumenten weer.

Tabel 4.13 Multitrait-multimethod matrix: correlaties tussen student- en tutoroordelen (N=176) berekend over gegevens in het academiejaar 1984/1985.

Dimensie	Student 1	2	3	4	5	6	7	Tutor 1	2	3	4	5	6	7
Student.														
1 Studielast	(75)													
2 Tutor-rol	01	(88)												
3 Ow-groep	18	38	(91)											
4 Taken	29	26	41	(79)										
5 Tutor-vakk.	03	26	03	01	(--)									
6 Onafh. blok	-17	-14	-33	04	00	(45)								
7 Variëteit	39	32	31	45	12	-16	(82)							
Tutor.														
1 Studielast	44	02	20	14	-01	-04	27	(52)						
2 Tutor-rol	-07	20	12	14	02	-22	07	15	(76)					
3 Ow-groep	17	19	62	30	-10	-26	15	19	18	(88)				
4 Taken	09	08	24	46	-03	-24	16	14	42	41	(79)			
5 Tutor-vakk.	00	30	02	03	46	08	07	08	00	-14	-07	(--)		
6 Onafh.	-12	-04	04	14	-03	-08	-08	03	06	20	35	-20	(76)	
7 Variëteit	-12	-06	08	12	-12	-15	16	27	11	24	39	-10	12	(77)

Noot. Alle correlaties worden zonder decimalen gepresenteerd. Correlaties groter dan .18 zijn statistisch significant ($p < .05$).

Campbell en Fiske (1959) hanteren vier criteria voor de beoordeling van een multitrait-multimethod matrix.

1. De waarden in de validiteitsdiagonaal dienen significant van nul te verschillen en voldoende hoog zijn. Als dat het geval is, dan is aan de eis van convergerende validiteit voldaan.
2. De waarden in de validiteitsdiagonaal moeten hoger zijn dan de overeenkomstige kolom en rij in de heterotrait-heteromethod driehoek. Als dit niet het geval is betekent dat, dat de convergerende validiteit op een trait niet onafhankelijk is van overeenstemming op andere traits.
3. De waarden in de validiteitsdiagonaal moeten hoger zijn dan de corresponderende correlaties tussen de betrokken trait en andere traits binnen een methode (de heterotrait-monomethod driehoek). Als deze correlaties de waarde van de betrouwbaarheidscoëfficiënt van de betreffende trait benaderen, kan er sprake zijn van methode- of halo-effecten.
4. In alle heterotrait-driehoeken (zowel voor de monomethode als de heteromethode blokken) moet het correlatiepatroon tussen de traits hetzelfde zijn. Wordt aan de laatste drie eisen voldaan, dan voldoet het onderzochte meetinstrument

aan de eis van divergerende validiteit.

Het eerste criterium van Campbell en Fiske is een toets voor de convergerende validiteit van het meetinstrument. Overeenkomstig dit criterium moeten de correlaties in de validiteitsdiagonaal significant en substantieel afwijken van nul. Uit de resultaten in tabel 4.13 blijkt dat de validiteitscoëfficiënten variëren tussen -.08 en .62. Vijf van de zeven coëfficiënten voldoen bovengenoemd criterium voor convergerende validiteit. Eén coëfficiënt is bijna significant op vijfprocents-niveau. Eén coëfficiënt voldoet niet aan het gestelde criterium.
Het tweede criterium van Campbell en Fiske vereist dat iedere convergerende validiteitscoëfficiënt een hogere waarde heeft dan ieder andere correlatiecoëfficiënt in dezelfde rij of kolom van de heterotrait-heteromethod-driehoeken. Deze toets vereist dat ieder van de zeven validiteitscoëfficiënten vergeleken wordt met twaalf andere coëfficiënten. In totaal vinden dus 84 vergelijkingen plaats. In 9 van de 84 gevallen wordt niet aan dit criterium voldaan. Zeven van die negen blijken veroorzaakt te worden door de niet significante waarde van de validiteitscoëfficiënt van de schaal 'onafhankelijk van het blokboek en onderwijsgroep studeren'.
Volgens het derde criterium moet iedere convergerende validiteitscoëfficiënt een hogere waarde hebben dan de correlaties tussen de corresponderende trait en de andere traits gemeten binnen een methode. Zo is de convergerende validiteitscoëfficiënt voor de schaal Onderwijsgroep 1) hoger dan de correlaties tussen de tutoroordelen op de schaal Onderwijsgroep en de tutoroordelen op de andere schalen, en 2) hoger dan de correlaties tussen de studentoordelen op de schaal onderwijsgroep en de studentoordelen op de andere schalen. Aan dit criterium werd 12 keer (uit 42 vergelijkingen) niet voldaan voor de studentschalen. Bij de tutorschalen voldeden 8 van de 42 vergelijkingen niet aan dit criterium. De schalen 6 (onafhankelijk van het blokboek en onderwijsgroep studeren) en 7 (varieteit in taakvormen en onderwerpen) waren voor het merendeel van deze verwerpingen verantwoordelijk. Deze resultaten suggereren dat er bij de studentoordelen enige sprake kan zijn van methode-effecten en/of van halo-effecten.

Volgens bovengenoemd vierde criterium zouden de correlaties tussen de schalen bij de studentoordelen een identiek patroon moeten hebben als de correlaties tussen de schalen bij de tutoroordelen. Inspectie van tabel 4.13 laat zien dat beide patronen redelijk overeenkomen. Ter nadere precisering werd de correlatie tussen de 21 buitendiagonaal-correlaties van de twee heterotrait-monomethod driehoeken berekend. Deze correlatie bedraagt .47 ($p<.05$). Dit resultaat wettigt de conclusie dat de correlatiepatronen redelijk goed overeenkomen. Aan dit criterium wordt dus in redelijke mate voldaan. De conclusie lijkt dan ook gerechtvaardigd dat een gedeelte van het veronderstelde methode-effect en/of halo-effect bij

studentoordelen (zie criterium 3) berust op ware correlaties tussen de schalen. Correlaties tussen schalen lijken dus onafhankelijk van de gevolgde methode te zijn.

4.7.8 Discussie en Conclusies.

Campbell en Fiske (1959) noemen vier criteria om de correlaties in een multitrait-multimethod-matrix te beoordelen. Aan de hand van deze criteria worden, op basis van correlaties tussen geobserveerde variabelen, conclusies getrokken over de onderliggende trait- en methodefactoren. Een probleem bij deze methode is dat Campbell en Fiske (1959) geen kwantitatieve normen geven waarmee bepaald kan worden of de onderzoeksresultaten voldoen aan de gestelde criteria. De interpretatie van de resultaten uit de in deze studie beschreven multitrait-multimethod analyse, wordt daardoor in zekere zin gecompliceerd. Zo blijft het de vraag of de correlaties tussen schalen werkelijke samenhangen representeren, dan wel onafhankelijk van elkaar zijn. De vraag is dus in welke mate dimensies van probleemgestuurd onderwijs echt met elkaar samenhangen. De statistisch significante correlatie bij studentoordelen tussen de schalen Taken en Onderwijsgroep (r=.41) kan bijvoorbeeld duiden op halo-effecten maar kan tevens de daadwerkelijke samenhang tussen beide schalen representeren. Theorieën over probleemgestuurd onderwijs moeten in dit geval de logische basis verschaffen waaraan correlaties tussen dimensies getoetst kunnen worden. Uit de resultaten in tabel 4.13 blijkt dat geen correlaties gevonden worden die in strijd zijn met de in hoofdstuk 2 beschreven theorieën over probleemgestuurd onderwijs.
Samenvattend kan ondanks het hierboven gesignaleerde probleem geconcludeerd worden dat de multitrait-multimethod-analyse een redelijk goede ondersteuning biedt voor de convergerende en divergerende validiteit van studentoordelen. In het onderzoek naar de convergerende validiteit werd gevonden dat de overeenstemming tussen studenten en tutoren op 5 van de 7 dimensies redelijk hoog is. In vergelijkbaar onderzoek worden voor convergerende validiteitscoëfficiënten ongeveer dezelfde waarden gevonden (Doyle & Crichton, 1978; Marsh, Overall & Kesler, 1979; Marsh, 1982). Aan de drie criteria voor de divergerende validiteit wordt eveneens redelijk goed voldaan. De correlatiepatronen tussen schalen binnen methoden vertonen voldoende overeenstemming en de correlaties tussen schalen lijken werkelijke samenhangen te representeren. Uit het gegeven dat zowel aan de eis van convergerende validiteit als aan de eis van divergerende validiteit redelijk goed wordt voldaan, kan dus geconcludeerd worden dat het in deze studie gebruikte instrument ter meting van studentoordelen voldoende constructvaliditeit bezit.

4.8 Studie 6: De invariantie van correlatiematrices.

4.8.1 Inleiding.

In studie 1 werd met behulp van exploratieve factoranalyse de onderliggende structuur van de beoordelingsvragenlijst voor studenten onderzocht. De uitkomsten van deze analyse werden gebruikt om schalen te construeren ten behoeve van de daaropvolgende studies. Studie 2 en 3 zijn betrouwbaarheidsstudies waarin de betrouwbaarheid van schalen en studentoordelen onderzocht is. In studie 2 werd de alpha-betrouwbaarheid van deze schalen berekend en in studie 3 werd de betrouwbaarheid van studentoordelen op deze schalen onderzocht. Studie 4 en 5 bestaan uit validiteitsstudies. In studie 4 werd de criteriumvaliditeit van studentoordelen onderzocht. In studie 5 was de constructvaliditeit van studentoordelen object van onderzoek. In laatstgenoemde studie werd aan de hand van een correlatiematrix waarin correlaties tussen schalen en methoden opgenomen waren, de divergerende en convergerende validiteit van studentoordelen onderzocht. In de nu volgende en tevens laatste studie, vormen drie correlatiematrices waarin correlaties tussen studentschalen opgenomen zijn, het uitgangspunt voor verder onderzoek naar de constructvaliditeit van studentoordelen.
Correlatiematrices kunnen als basis voor factoranalyse dienen. Aan de hand van de daaruit resulterende factoren kunnen conclusies getrokken worden over de onderliggende (latente) dimensies die aan de geobserveerde variabelen ten grondslag liggen. De vraag die zich steeds weer voordoet, is in hoeverre samenhangen binnen een multidimensionele structuur een accurate weerspiegeling vormen van de samenhangen tussen de dimensies in de werkelijkheid. In de vorige studie is deze vraag onderzocht door correlaties tussen dimensies en methoden te analyseren. In studie 6 wordt nadere beantwoording van deze vraag gezocht door vergelijking van correlaties tussen dimensies van studentoordelen, over drie opeenvolgende academiejaren. Uitgangspunt in deze studie is dat correlaties tussen dimensies van probleemgestuurd onderwijs constant zijn en dus onafhankelijk van het academiejaar waarin de metingen plaatsvonden. Dit veronderstelt tevens dat gedurende de onderzoeksperiode geen essentiële veranderingen in de opzet van probleemgestuurd onderwijs optraden. Volgens deze uitgangspunten bezitten studentoordelen voldoende constructvaliditeit als de correlatiematrices over een aantal academiejaren invariant zijn (Marsh & Hocevar 1984).

4.8.2 Procedure.

Het onderzoek had betrekking op de academiejaren 1984/1985, 1983/1984 en 1982/1983. Groepsoordelen werden gebruikt om correlaties tussen de in studie 1 geïdentificeerde schalen te berekenen. Per academiejaar werd een correlatiematrix opgesteld die de volgende schalen bevatte:

- schaal 1 Kwaliteit van de taken
- schaal 2 Tutor
- schaal 3 Onderwijsgroep
- schaal 5 Moeilijkheidsgraad
- schaal 6 Bloktoets
- schaal 7 Leermiddelen
- schaal 8 Globaal oordeel
- schaal 10 Inhoudsdeskundigen
- schaal 12 Onafhankelijke studie
- schaal 13 Structurering Onderwijs
- schaal 14 Afwisseling
- schaal 15 Tutor hard werken

In bijlage 3 worden deze matrices weergegeven. Schaal 4 en 9 werden niet in het onderzoek betrokken vanwege hun lagere alpha-betrouwbaarheid. Gegevens uit het academiejaar 1981/1982 werden niet gebruikt vanwege het ontbreken van de benodigde schaalgegevens.

4.8.3 Methode.

Met behulp van het programma LISREL IV werden de correlatiematrices op hun invariantie onderzocht (Jöreskog & Sörbom, 1978). Met dit programma kan men covarianties of correlaties tussen variabelen analyseren in termen van een causale structuur tussen die variabelen. Het programma toetst in hoeverre de correlaties tussen geobserveerde variabelen kunnen worden gereproduceerd door een a priori gespecificeerd causaal model.
In deze studie wordt de LISREL-notatie gebruikt om modellen te specificeren. De in de steekproef geobserveerde matrix noemt men S, de populatie correlatiematrix wordt SIGMA genoemd. Met behulp van LISREL kan men nagaan hoe groot de kans is dat S uit een populatie voortkomt, waarin de relaties zijn zoals het model ze specificeert en tot uiting laat komen in SIGMA. Is deze kans klein, dan neemt men aan dat het verschil tussen SIGMA en S niet het gevolg is van steekproeffouten maar van fouten in de specificatie van het model. Het model wordt in dat geval verworpen.
Het model dat in deze analyse gebruikt wordt veronderstelt de specificatie van drie matrices. Deze zijn achtereenvolgens de LAMBDA-X matrix die de factorladingen bevat, de PHI matrix die de correlaties tussen factoren bevat, en de THETA-DELTA matrix die de unieke varianties ofwel de meetfouten van iedere geobserveerde variabele bevat. Op basis van deze matrices wordt volgens een a priori gepostuleerd model een correlatiematrix SIGMA (xx) gereproduceerd. De gereproduceerde correlatiematrix wordt met behulp van matrixnotatie gedefinieerd als:

(1) $$\Sigma_{xx} = \Lambda_x \Phi \Lambda_x' + \Theta_\delta$$

In deze studie wordt onderzocht of het volgende geldt:

(2) $$\Sigma_1 = \Sigma_2 = \Sigma_3$$

Σ_1 = de gereproduceerde correlatiematrix uit het academiejaar 1982/1983,
Σ_2 = de gereproduceerde correlatiematrix uit het academiejaar 1983/1984,
Σ_3 = de gereproduceerde correlatiematrix uit het academiejaar 1984/1985.

Om de gelijkheid van deze correlatiematrixen te toetsen moeten de volgende specificaties in vergelijking (1) gemaakt worden:

$$\Lambda_x^g = I \quad \text{en} \quad \theta_\delta^g = 0$$

voor alle 3 groepen g (g=1,..3). Daardoor wordt $SIGMA_{xx}$ in het model gelijk aan PHI. Bovendien geldt dan dat S=SIGMA en SIGMA(1)=SIGMA(2)=SIGMA(3) hetgeen gelijk is aan S(1)=S(2)=S(3). Om vergelijking (2) te onderzoeken wordt getoetst of de volgende hypothese waar is:

$$\Phi_1 = \Phi_2 = \Phi_3$$

PHI is de correlatiematrix in de populatie en g (1,..3) is de index voor de betreffende populatie.

4.8.4 Resultaten.

Om een uitgangspunt te hebben voor de modelfit werd eerst onderzocht of de buitendiagonaal-correlaties van de originele correlatiematrices substantieel waren (groter dan nul). Als nulmodel werd gekozen voor een situatie waarbij alle correlaties buiten de diagonaal nul zijn. Een CHI-kwadraat toets leverde het volgende resultaat op: CHI-kwadraat = 3024.52, df = 234, p = 0.00.
In deze studie wordt getoetst of de correlatiematrices invariant zijn. De vraag is dus of alle elementen uit deze matrices gelijk zijn. Geobserveerde verschillen zijn dan random fluctuaties. De toets van gelijke PHI - matrixen leverde de volgende resultaten op: CHI-kwadraat = 262.97, df = 156, p = 0.00.
Het LISREL programma verschaft een CHI-kwadraat toets waarin de originele matrix met de volgens een model gereproduceerde matrix vergeleken wordt. CHI-kwadraat is een maat voor de 'goodness of fit'.
Een significante CHI-kwadraat is een teken voor verschillen tussen de originele en gereproduceerde matrix. Deze toets is echter zoals alle toetsen afhankelijk van de steekproefomvang. Een goede fit kan bij grote populaties toch tot een significante CHI-kwadraat leiden.
Andersom kan bij een slechte fit een nonsignificante CHI--

kwadraat verkregen worden als de steekproef klein is. Alternatieve indices om de modelfit te beschrijven zijn de ratio van CHI-kwadraat, het aantal vrijheidsgraden en de maten van Bentler en Bonett (1980), namelijk: rho en delta. Als de waarden van delta en rho hoger zijn dan .90 is er weinig tot geen verschil tussen de originele en gereproduceerde correlatiematrixen. In tabel 4.14 zijn de resultaten van beide analyses weergegeven.

Tabel 4.14 Samenvatting van de getoetste modellen.

	CHI^2	df	p	rho	delta
nulmodel	3024.52	234	0.00	-	-
model H(Sigma)	262.97	156	0.00	.94	.91
Verschil	2761.55	78	0.00	-	-

Het model H(Sigma) heeft een CHI-kwadraat van 1.69 per vrijheidsgraad, hetgeen niet significant is. Het verschil met het nulmodel is wel significant. De waarden van zowel Rho als Delta zijn hoger dan de door Bentler en Bonnet (1980) gestelde kritische grens van .90. Deze resultaten suggereren dat het model H(SIGMA) een correlatiematrix reproduceert die niet significant afwijkt van de geobserveerde correlatiematrices. Daaruit kan geconcludeerd worden dat de geobserveerde correlatiematrices met betrekking tot de academiejaren 1982/1983, 1983/1984 en 1984/1985 invariant zijn. Correlaties tussen dimensies zijn dus met andere woorden invariant en onafhankelijk van het academiejaar waarin de metingen plaatsvonden.

4.8.5 Discussie en conclusies.

Het doel van deze studie was om de multidimensionaliteit van studentoordelen te onderzoeken. In studie 1 en studie 5 was dit reeds gebeurd met meer 'klassieke' technieken: exploratieve factoranalyse (principale componenten analyse) en multitrait-multimethod analyse.. In deze studie werd met behulp van confirmerende factoranalyse de multidimensionaliteit van studentoordelen onderzocht. Correlaties tussen dimensies binnen een academiejaar vormden het uitgangspunt voor statistische analyses met behulp van LISREL IV. Voor drie opeenvolgende academiejaren werd een correlatiematrix opgesteld die de correlaties tussen dimensies bevatte. In deze studie werd aangetoond dat de betreffende correlatiematrices een multivariate structuur bezitten die invariant is over de onderzochte drie academiejaren. Onder de assumptie dat in de onderzoeksperiode geen wezenlijke veranderingen in de opzet van probleemgestuurd onderwijs plaatsvonden, kunnen de correlaties tussen dimensies dus opgevat worden als ware correlaties. Methode- of halo-effecten kunnen derhalve nauwelijks als verklarende factor gezien worden voor de correla-

ties tussen dimensies. Samenvattend mogen we concluderen dat de resultaten uit deze studie een sterke indicatie vormen voor de constructvaliditeit van de in dit proefschrift beschreven vragenlijst.

4.9 Conclusies.

In dit hoofdstuk zijn zes studies naar de betrouwbaarheid en validiteit van studentoordelen beschreven. De voornaamste resultaten van deze studies kunnen als volgt worden samengevat.
In studie 1 bleek dat de beoordelingsvragenlijst voor studenten een multidimensionele structuur heeft die bovendien redelijk goed overeenkomt met de hypothetische structuur. Daaruit kan geconcludeerd worden dat studenten blijkbaar een aantal beoordelingsdimensies hanteren die overeenkomen met de door de onderzoekers veronderstelde dimensies. Dit resultaat is een eerste indicatie voor de validiteit van de vragenlijst. Bovendien blijkt uit de principale-componentenanalyse dat het optreden van systematische meetfouten, waaronder halo-effecten, weinig aannemelijk is.
In studie 2 kwam naar voren dat op basis van de uitkomsten van de principale-componentenanalyse schalen geconstrueerd kunnen worden die voor het merendeel intern consistent zijn.
Uit studie 3 bleek dat individuele studentoordelen en oordelen van onderwijsgroepen een hoge betrouwbaarheid bezitten. De gevonden betrouwbaarheden van individuele studentoordelen zijn over het algemeen iets hoger dan in literatuur gerapporteerd wordt (Feldman, 1977; Marsh & Overall, 1979). Het blijkt dat het middelen van individuele studentoordelen binnen onderwijsgroepen de betrouwbaarheid van studentoordelen, over elementen van het blok die voor alle onderwijsgroepen gelijk zijn, aanzienlijk verhoogt.
In studie 4 werd de criteriumvaliditeit van studentoordelen onderzocht. Studentoordelen werden in deze studie gecorreleerd met bloktoetsresultaten. De resultaten die in deze studie gevonden worden zijn op zichzelf verrassend en onverwacht. In de literatuur worden namelijk in het merendeel van de onderzoeken redelijk positieve correlaties tussen studentoordelen en tentamenresultaten gevonden (Cohen, 1981; Dowell & Neal, 1982). In studie 4 werden daarentegen wisselende correlaties gevonden. In sommige academiejaren was wel sprake van redelijk positieve correlaties, in andere jaren niet. Een verklaring voor deze resultaten werd gezocht in de wisselende status (formatief of summatief) van de bloktoets bij beslissingen over studievoortgang. Een voor de hand liggende vraag was dan ook of deze resultaten de validiteit van studentoordelen juist wel of niet indiceren. In studie 5 werd de validiteitsvraag onderzocht met behulp van multitrait-multimethod analyse. Uit deze studie kwam naar voren studentoordelen een redelijk goede constructievaliditeit bezitten. In dit geval was het validatiecriterium de convergerende en divergerende validiteit van studentoordelen zoals gerelateerd aan tutor-

oordelen. De resultaten uit deze studie geven een verdere aanwijzing voor de in studie 4 gesignaleerde problematiek. De mogelijkheid bestaat immers dat het in studie 4 gebruikte criterium, namelijk studieprestaties, geen adequaat extern criterium vormt waaraan de beoordelingsvragenlijst voor studenten gevalideerd kan worden. Deze hypothese wordt verder gesteund door de resultaten uit studie 6.
In studie 6 bleek dat correlaties tussen schalen van de vragenlijst onafhankelijk zijn van het academiejaar waarin de metingen plaatsvinden. Met andere woorden, de correlatiematrices van de afzonderlijke academiejaren blijken een invariante structuur te bezitten. Uit dit resultaat kan geconcludeerd worden dat de correlaties tussen de onderliggende dimensies van de vragenlijst stabiel zijn over een bepaalde tijdsperiode (in dit geval drie academiejaren), hetgeen lijkt te wijzen op een redelijk goede constructvaliditeit van studentoordelen. Vergelijkbare resultaten worden ook door Marsh en Hocevar (1983, 1984) gerapporteerd.
De uitkomsten uit de studies naar de betrouwbaarheid en validiteit van studentoordelen lijken derhalve de conclusie te rechtvaardigen, dat de in dit proefschrift beschreven vragenlijst een betrouwbaar en valide instrument is om de kwaliteit van probleemgestuurd medisch onderwijs te meten.

HOOFDSTUK 5. STUDIES NAAR DE BRUIKBAARHEID VAN STUDENTOORDELEN.

5.1 Inleiding.

Het evaluatiesysteem van de medische faculteit heeft, zoals toegelicht in hoofdstuk 2, de functie belanghebbenden (bestuurders en docenten) van zodanige informatie te voorzien dat de kwaliteit van het onderwijs beheerst en verbeterd kan worden. Twee factoren spelen een rol bij de realisatie van deze doelstelling: 1) de beschikbaarheid van betrouwbare en valide gegevens en 2) inzicht in de wijze waarop die gegevens ten behoeve van de onderwijsorganisatie gebruikt (kunnen) worden.
In de vorige hoofdstukken is voornamelijk aandacht besteed aan het eerste aspect, namelijk de meetbaarheid van de kwaliteit van onderwijsprogramma's en de betrouwbaarheid en validiteit van het gehanteerde instrument. In dit hoofdstuk zullen we ons wijden aan het tweede aspect: het gebruik van evaluatiegegevens.
In dit hoofdstuk wordt eerst kort teruggeblikt naar een gedeelte van hoofdstuk 1, wat betreft een aantal theorieën en onderzoeken met betrekking tot het gebruik van evaluatieresultaten. De betreffende paragraaf geeft een korte beschrijving van de problemen die een effectief gebruik van evaluatieresultaten met zich meebrengt. In de daaropvolgende paragrafen wordt aan vier voorbeelden geïllustreerd hoe evaluatiegegevens gebruikt zijn en welke rol zij spelen in het onderwijsbeleid van de faculteit. In de beschrijving van die voorbeelden staan twee onderzoeksvragen centraal: "Hoe is de informatie die evaluatieactiviteiten opleveren, gebruikt bij besluitvorming over het onderwijs? en "Welke effecten had deze informatie op de kwaliteit van het onderwijs?"
Eerst wordt aan de hand van twee case-studies geïllustreerd hoe twee planningsgroepen van evaluatiegegevens gebruik gemaakt hebben en tot welk resultaat dit leidde. In deze studies wordt beschreven op welke wijze de betreffende planningsgroepen een aantal problemen in hun blok aangepakt hebben en welk resultaat deze interventies hadden op de kwaliteit van het betreffende onderwijs. Daarna wordt een studie beschreven waarin het gebruik van informatie op curriculumniveau centraal staat. Deze studie heeft een beleidsgericht karakter. Gerapporteerd wordt hoe de onderwijscommissie van de medische faculteit experimenteerde met de werkwijze en samenstelling van planningsgroepen. Dit experiment beoogde de kwaliteit van het onderwijs te verbeteren door tutoren mede verantwoordelijk te maken voor de voorbereiding van een blok, voor het samenstellen van het blokboek en voor het formuleren van toetsitems. Met behulp van de evaluatieaanpak die hier beschreven is, werd dit experiment geëvalueerd en mede op basis van die gegevens nam de onderwijscommissie beslissingen

over de continuering van het experiment. De vierde studie heeft betrekking op het gebruik van evaluatiegegevens door tutoren. In deze studie staat de vraag centraal of feedback omtrent het functioneren van tutoren effect heeft op het gedrag van die tutoren tijdens de vervulling van daaropvolgende tutorschappen.

5.2 Het gebruik van evaluatieresultaten: een korte terugblik naar hoofdstuk 1.

Onderwijsevaluatie vervult binnen de medische faculteit een formatieve functie. Het faculteitsbestuur, de onderwijscommissie, de planningsgroepen en de tutoren kunnen de evaluatiegegevens gebruiken om de kwaliteit van het onderwijs te verbeteren maar zijn daartoe niet verplicht. De vraag die zich dan ook opdringt, is of ook daadwerkelijk gebruik gemaakt wordt van deze gegevens. In hoofdstuk 1 werden vijf factoren genoemd die volgens Levinton en Hughes (1981) het gebruik van evaluatieresultaten beïnvloeden: de relevantie van de informatie, de wijze waarop de informatie wordt doorgespeeld, de verwerking van de informatie, de geloofwaardigheid van de informatie en de betrokkenheid van de gebruikers bij de evaluatie. Bovendien werden in hoofdstuk 1 een aantal onderzoeken beschreven naar de invloed van studentoordelen op de kwaliteit van het onderwijs. In deze onderzoeken stond de vraag centraal onder welke condities terugkoppeling van informatie, verkregen met behulp van studentoordelen, een positief effect had op de kwaliteit van het onderwijs in cursussen en op het doceergedrag van docenten.
In tabel 5.1 worden voor alle duidelijkheid de door Levinton & Hughes (1981) genoemde factoren nog eens gepresenteerd. In de studies die in de volgende paragrafen gepresenteerd worden, zal regelmatig naar deze factoren verwezen worden.
Een ander belangrijk punt in hoofdstuk 1 was de vraag hoe "gebruik van evaluatieresultaten" geïnterpreteerd moest worden. Deze vraag was van belang voor het formuleren van criteria waarmee feitelijk gebruik van evaluatieresultaten aangetoond zou kunnen worden. Met name wat betreft: instrumenteel en conceptueel gebruik.
Instrumenteel gebruik verwijst naar expliciet gebruik van evaluatieresultaten bij het nemen van beslissingen. Conceptueel gebruik verwijst naar het beïnvloeden van de gedachtengang van personen die beslissingen nemen, zonder dat een beslissing een direct resultaat is van de gepresenteerde informatie.

Tabel 5.1: Factoren die het gebruik van informatie beïnvloeden

1. RELEVANTIE.
   - aansluiting van de evaluatie op de behoeften van de gebruiker,
   - tijdsplanning van het evaluatie-onderzoek: tijdstip van presentatie.

2. COMMUNICATIE.
   - directheid van de communicatie tussen gebruiker en evaluator,
   - verspreiding van de informatie.

3. PRESENTATIE.
   - presentatie van informatie,
   - mate van begrip van de gepresenteerde informatie bij de gebruiker.

4. GELOOFWAARDIGHEID.
   - overeenstemming met andere informatiebronnen,
   - vooropgezette meningen van de gebruiker ten aanzien van de bruikbaarheid van het onderzoek in het algemeen,
   - geloofwaardigheid van de evaluator,
   - kwaliteit van het onderzoek.

5. BETROKKENHEID VAN DE GEBRUIKER.
   - persoonlijke betrokkenheid met het evaluatie-onderzoek.

5.3 Studie 1: case blok 3.3 "Pijn op de borst".

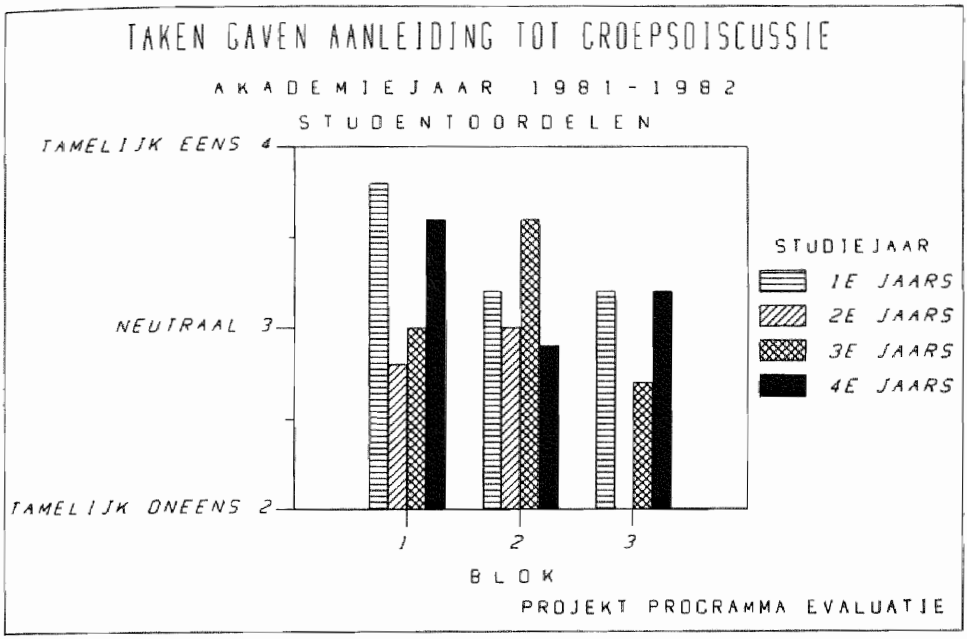
De eerste studie beschrijft een case die laat zien hoe concrete verbeteringen in een onderwijsblok tot stand kwamen en hoe dit merkbaar was in de evaluatie-uitkomsten van het blok. Het betreffende blok wordt in deze casestudie over een periode van vier academiejaren gevolgd. Gedurende deze periode werden een aantal ingrijpende wijzigingen in de opzet van het blokboek doorgevoerd.
Aanleiding tot deze wijzigingen waren een aantal problemen die in het begin van deze periode uit de onderwijsevaluatie naar voren kwamen. Het blok ontwikkelde zich van een relatief laag, dat wil zeggen slecht beoordeeld blok, tot het best beoordeelde blok van het onderwijsprogramma in de eerste vier studiejaren. Deze casestudie is om twee redenen van belang. In de eerste plaats wordt het belang getoond van een goede communicatie (met name de kwaliteit van de communicatie en minder de kwantiteit) tussen betrokkenen. Een belangrijk gegeven in deze studie is dat gedurende het betreffende tijdvak het coördinatorschap van de planningsgroep in handen bleef van dezelfde persoon. In de tweede plaats toont deze studie hoe het evaluatie-instrument bruikbaar is voor het meten van kwaliteitsverbeteringen in het onderwijs.

5.3.1 Probleembeschrijving blok 3.3.

Blok 3.3 (het derde blok in het derde studiejaar) heeft als titel "Pijn op de borst". Het bestaat uit een inleiding in de cardiologie en pulmonologie. Het blok is onderverdeeld in twee subthema's die ieder drie weken in beslag nemen. De subthema's hebben respectievelijk betrekking op klachten op het gebied van het hart en dat van de longen.
In het academiejaar 1981/1982 werd blok 3.3 door de evaluatoren gekwalificeerd als een problematisch blok. Uit de evaluatie bleek namelijk dat het blok in zijn geheel een lagere beoordeling kreeg dan alle andere blokken die op dat moment geëvalueerd waren. Uit de evaluatiegegevens kwam naar voren dat met name de taken, het blokboek en het functioneren van de onderwijsgroepen relatief slecht beoordeeld werden. Het blokboek bevatte dat jaar acht casus betreffende hart- en vaatziekten en zeven casus betreffende longziekten.
De bron van de problemen werd volgens de studenten gevormd door de opzet van het blokboek en de vorm van de taken. Volgens de studenten waren de taken te gestructureerd, werd de groepsdiscussie te weinig gestimuleerd door de taken en gaven de taken relatief weinig aanleiding tot zelfstudie. Uit de antwoorden op de open vragen kwam onder andere naar voren dat men de casus te klassiek vond: "Een blokboek is geen leerboek", "De casus geven te weinig inzicht in wat je in de praktijk kunt tegenkomen". Een ander punt van kritiek was de vormgeving van de taken. In het blokboek was alle relevante

informatie over de casus opgenomen. Studenten hadden liever gezien dat deze informatie niet in het blokboek vermeld stond, maar dat de tutor over de informatie beschikte. De casus gaf volgens de studenten door haar uitgebreidheid geen aanleiding tot discussie in de onderwijsgroep noch tot zelfstudie van studenten. De casus hadden namelijk een structuur waarin eerst een patiëntgeval beschreven werd, waarna een aantal vragen volgden. Tenslotte werden op de volgende pagina('s) de antwoorden op de eerder gestelde vragen gegeven.
Het blokboek kreeg daardoor het karakter van een studieboek waarin door middel van een vraag-aanbod vorm de leerstof behandeld werd. Een bezwaar van deze aanpak was dat de leerdoelen in de casus reeds vooraf door de planningsgroep behandeld werden. Met als gevolg dat studenten de noodzaak niet voelden om de casus op probleemgestuurde wijze te analyseren, noch om zelfstandig naar verklaringen te zoeken voor de aangeboden problemen. De casus genereerde als het ware te weinig vragen die in de onderwijsgroep bediscussiëerd konden worden en een basis vormden voor zelfstudie. Het functioneren van de onderwijsgroepen vormde om die reden een probleem in het blok. De groepen maakten minder gebruik van systematische werkprocedures (Zevensprong, SOEP; zie hfd 2), de groep stimuleerde minder tot zelfstudie-aktiviteiten, de bijeenkomsten waren minder productief, niet iedereen leverde een aktieve bijdrage en men hield zich matig tot slecht aan de gemaakte afspraken.
De hierboven gesignaleerde problemen kwamen in de evaluatie van het blok zowel naar voren in de standaard items over het blokboek, de taken en het functioneren van de onderwijsgroep als in de open vragen van de vragenlijst. Als voorbeeld van de relatief lage beoordeling van het blokboek en de taken geeft de gemiddelde scores op vraag 14 "De taken gaven voldoende aanleiding tot een zinnige groepsdiscussie" voor alle blokken (1-3) die op dat moment geëvalueerd waren. Uit figuur 5.1 blijkt dat, in vergelijking met de andere blokken, de taken relatief weinig aanleiding gaven tot een zinvolle groepsdiscussie.

Figuur 5.1 Studentoordelen vraag 14, blok 1-3, 1981/1982.

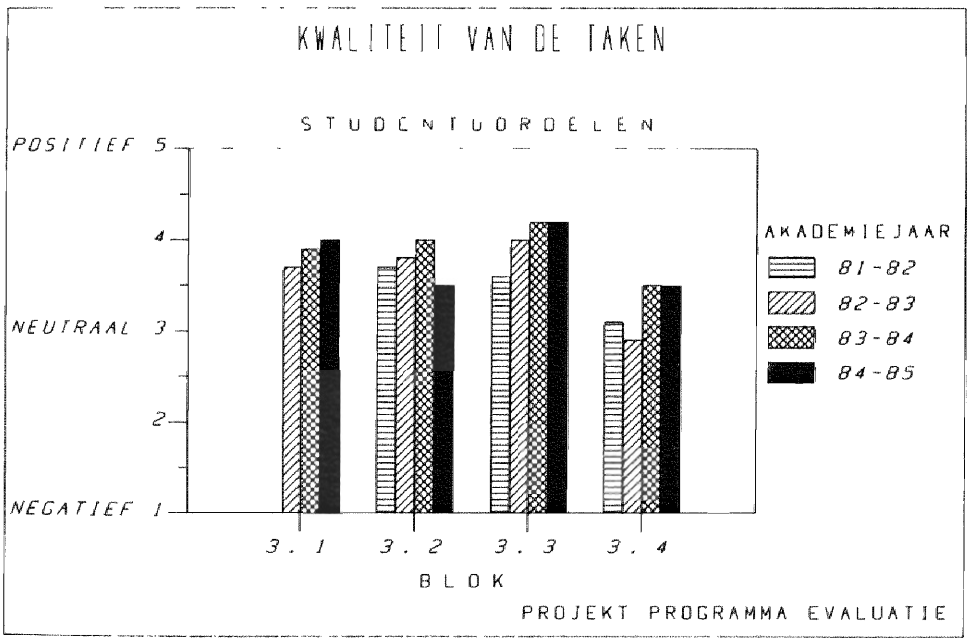


De meeste andere items over het blokboek en de taken vertoonde eenzelfde patroon: een gemiddelde score die in vergelijking met de andere blokken laag was. Een verklaring voor het slechte functioneren van de onderwijsgroepen en de lage beoordeling van de taken werd door de projectgroep programma-evaluatie gezocht in de vormgeving van de taken.
Blok 3.3 werd overigens niet op alle onderdelen van de standaardvragenlijst laag beoordeeld. De items van de onderdelen "Algemene Indruk" en "Tutor" werden redelijk positief gescoord. Met andere woorden, de problemen in het blok waren niet gecorreleerd aan kenmerken van de leerstof, aansluiting op voorkennis, eventuele onduidelijkheden over de blokdoelstellingen, een gebrek aan motivatie bij studenten of een slecht functionerende groep tutoren. Over andere aktiviteiten in het blok (practica, patiëntcontacten, bezoek aan Corony Care Unit) waren studenten zelfs zeer goed te spreken. Een uitzondering hierop werd gevormd door één bepaald practicum.
In het evaluatierapport van blok 3.3 werden bovengenoemde punten vermeld. Het rapport bevatte een commentaar op de evaluatieresultaten, geschreven door een lid van de projectgroep, plus de weergave van gemiddelde en standaarddeviatie per item. De planningsgroep ontving dit rapport ongeveer zes

weken na afloop van het blok.
Korte tijd later nam de blokcoördinator van de betreffende planningsgroep contact op met de auteur van het evaluatierapport met de vraag om toelichting. De centrale vraag van de coördinator was in hoeverre de opmerkingen uit het rapport betreffende het blokboek en het practicum representatief en valide waren. Hij betwijfelde of de kritische opmerkingen van studenten over het blokboek en het practicum, representatief waren voor de meningen van alle studenten die het blok gevolgd hadden dan wel afkomstig kon zijn van een kleine groep studenten die een wat kritischer instelling hadden ten opzichte van het blok. Met andere woorden, de planningsgroep vroeg zich af in hoeverre de resultaten serieus genomen moesten worden.
Aangezien deze vragen niet afdoende beantwoord konden worden, werden door de projectgroep drie aktiviteiten ondernomen: 1) aan een kleine steekproef (12) studenten werd een vragenlijst voorgelegd met betrekking tot de problematiek rondom het practicum; 2) een vijftal aselect gekozen studenten werd geïnterviewd over het blok en 3) de bloktoetsresultaten, het blokboek en de taken werden bestudeerd door twee leden van de projectgroep.
De resultaten uit de steekproefgewijze afgenomen vragenlijst over de gang van zaken in het practicum bevestigden de signalen uit de standaardevaluatie. Het betreffende practicum werd door de studenten als onbevredigend ervaren ("te weinig leerzaam"). Uit de interviews bleek opnieuw dat de onderwijsgroepen vóór de studenten nogal frusterend verlopen waren, omdat de casus geen aanleiding gaven tot het uitwisselen van ideeën. Volgens de geïnterviewde studenten waren de casus "te sterk gestructureerd" om een zinvolle basis voor discussie te bieden. Bestudering van het blokboek en de bloktoets gaven eveneens geen aanleiding tot wijziging van eerdere visies van de projectgroep op het blok.
Het tweede contact bestond uit een gesprek tussen twee leden van de projectgroep, de blokcoördinator en de practicumdocent. In dat gesprek werden de resultaten uit de, hierboven genoemde, aanvullende evaluatie-aktiviteiten besproken en werden adviezen gegeven voor veranderingen in de opzet van de taken. De suggestie werd gegeven om de vragen en antwoorden bij bepaalde taken weg te laten, om zodoende studenten te stimuleren zelf vragen te formuleren en zelf mogelijke antwoorden te bespreken in de onderwijsgroep. Een andere suggestie was om de vragen bij bepaalde taken te laten staan en de antwoorden aan de tutor te geven. Een laatste suggestie betrof de opbouw van het blok in twee subthema's. De overweging werd gegeven om het blok in drie delen van twee weken op te zetten: twee weken hart, twee weken longen en twee weken over kwesties waarin dat allemaal niet zo duidelijk ligt (gemengde problematiek rondom het thema "Pijn op de borst").
Na dit gesprek vonden in dat jaar geen verdere contacten meer

plaats tussen de planningsgroep en de projectgroep.
Het volgende academiejaar, namelijk 1982/1983, bleek dat blok 3.3 over het geheel genomen aanmerkelijk beter beoordeeld werd dan het jaar daarvoor. Het blokboek, de taken en het functioneren van de onderwijsgroepen werden positief beoordeeld. Het daaropvolgende academiejaar (1983/1984) zette de stijgende trend door. Dit had uiteindelijk tot gevolg dat blok 3.3 het best beoordeelde blok van het derde studiejaar werd, en zelfs het best beoordeelde blok van het onderwijsprogramma in de eerste vier studiejaren. Het jaar daarna (1984/1985) kwam de stijging tot stilstand en bleef het blok het best beoordeelde blok van het programma. In figuur 5.2 zijn de scores weergegeven van de takenschaal (uitgedrukt in een gemiddelde score over de items waaruit die schaal bestaat). Uit de figuur wordt duidelijk dat de scores toenamen van 3.6, op een schaal van 1 tot en met 5, naar 4.2. De grootste stijging in de beoordeling doet zich voor in het academiejaar 1982/1983 (het jaar waar het gesprek tussen evaluatoren en vertegenwoordigers van de planningsgroep plaatsvond en waarin de meeste veranderingen in de opzet van het blokboek plaatsvonden). Het daaropvolgende jaar werden, blijkens een analyse van het blokboek, naar verhouding minder veranderingen doorgevoerd. Dit wordt weerspiegeld in de afgenomen stijging. In het academiejaar 1984/1985 was het blokboek identiek aan het jaar daarvoor. Uit figuur 5.2 blijkt dat in dat jaar de beoordeling van de taken in het blokboek identiek is aan die van het jaar daarvoor.

Figuur 5.2 Studentoordelen over taken in 4 derdejaarsblokken van 1981/1982 tot en met 1984/1985.



Tabel 5.2: Scores op de takenschaal: gemiddelde, standaard-deviatie, en F-toets.

Blok 3.3 Academiejaar	Gem.	SD	N	F-ratio	Df	P
1981/1982	3.6	.4	10	15.81	(3,51)	.0(
1982/1983	4.0	.3	12			
1983/1984	4.2	.2	15			
1984/1985	4.2	.2	18			

Samenvattend toont de figuur duidelijk hoe over een periode van vier academiejaren, blok 3.3 zich ontwikkelde van een relatief slecht beoordeeld blok tot een relatief goed beoordeeld blok (mede gezien de vergelijkbare stijging op schalen betreffende het functioneren van de onderwijsgroep).

Een voor de hand liggende vraag is welke veranderingen in blok 3.3 aanleiding gaven tot de gestegen waardering voor het blok. Een andere vraag is in hoeverre de evaluatie van het blok een rol heeft gespeeld bij die veranderingen; met andere woorden, hoe zijn de evaluatieresultaten gebruikt en welk effect had de evaluatie op de kwaliteit van het onderwijs in het blok.
In het laatste academiejaar waarover hier gerapporteerd wordt, 1984/1985, vond, nadat het evaluatierapport van blok 3.3 naar de planningsgroep was gestuurd, een gesprek plaats tussen de blokcoördinator en de evaluator. In dat gesprek werd onder andere aandacht besteed aan de sterk toegenomen waardering voor het blok. Een verklaring voor dit fenomeen werd door beiden gezocht in de veranderde opzet van het blokboek, zoals die in 1982/1983 vorm had gekregen. Zoals gezegd had het blokboek in het academiejaar 1981/1982 een opzet die te gestructureerd was, weinig aanleiding gaf tot discussie, etc. Het jaar daarna werd de vorm van de casus ingrijpend veranderd; de onderwerpen die de in casus behandeld werden, bleven evenwel hetzelfde. Alle casus uit het gedeelte "Longen" en enkele casus uit het gedeelte "Hart", bevatten in de nieuwe opzet een korte situatie- of klachtbeschrijving waarna een aantal vragen gesteld werden. De tutor had de antwoorden op deze vragen. Het was de taak van de studenten om systematisch de casus te analyseren, door gericht vragen te stellen, hypotheses te formuleren, verklaringen te bedenken en oplossingen te formuleren. De tutor verschafte de studenten desgewenst informatie, over de gestelde vragen. In onderstaande tabel is als voorbeeld een casus opgenomen in de oude en nieuwe opzet.

Tabel 5.3: Voorbeeld karakteristieke casus in blok 3.3 1981/1982 en dezelfde maar gewijzigde casus in 1982/1983.

CASUS 1 1981/1982. *BLZ. 1*

Jan van Helden, 18 jaar oud, wordt 's morgens om 6 uur wakker met pijn op de borst. De pijn zit voor op de borst, is schurend van karakter en wordt erger wanneer Jan op zijn zij gaat liggen of moet hoesten. Hij merkt dat hij nog de minste last van pijn heeft wanneer hij rechtop gaat zitten. De pijn is zo erg dat de moeder van Jan meteen de dokter laat komen.

- *Wat denkt u dat hier aan de hand is?*
- *Welke specifieke vragen gaat u stellen bij de anamnese?*
- *Waar gaat u op letten bij het lichamelijk onderzoek?*
- *Laat u de patiënt opnemen in het ziekenhuis?*

Beantwoord deze vragen voordat u verder gaat met de volgende bladzijde.

CASUS 1 1981/1982. *BLZ. 2*

Bij lichamelijk onderzoek is de bloeddruk 120/80 mmHg. De pols is 100/min., regulier en aequaal. Aan hoofd en hals worden geen afwijkingen gevonden. Er zijn geen abnormale precordiale pulsaties.
De ictus cordis is zonder bijzonderheden.

Bij auscultatie van het hart is er een graad 4/6 ruw systolisch-diastolisch geruis te horen.

Tijdens het onderzoek wil Jan voortdurend rechtop gaan zitten. Hij heeft veel last van pijn, vooral wanneer hij achterover ligt of op zijn zij. De temperatuur blijkt 38.5 graden C te bedragen.

De huisarts denkt aan een bepaalde diagnose en laat Jan opnemen in het ziekenhuis.

- *Wat is uw diagnose?*
- *Wat denkt u dat er in het ziekenhuis aan verder onderzoek zal gebeuren?*

Beantwoord eerst deze vragen en ga dan verder met de volgende bladzijde.

CASUS 1 1981/1982. *BLZ. 3*

Bij opname in het ziekenhuis bevestigt de cardioloog de bevindingen van de huisarts. Hij maakt een cardiogram waarop ST-segment elevaties te zien zijn in alle afleidingen.
Op de thoraxfoto is het hart niet vergroot. Er bevindt zich een kleine hoeveelheid vocht in de linker sinus pleurea.
Bij bloedonderzoek worden behalve een bezinking van 30 mm in het eerste uur, geen afwijkingen gevonden.

Jan krijgt bedrust voorgeschreven en aspirine.

- *Waarop denkt u moet vooral gelet worden tijdens het verblijf van Jan in het ziekenhuis en hoe lang denkt u dat hij in het ziekenhuis moet blijven?*
- *Wat zijn de risico's na ontslag uit het ziekenhuis?*

De tutor heeft de antwoorden op deze vragen.

CASUS 1 1982/1983. *BLZ. 1*

Jan van Helden, 18 jaar oud, wordt 's morgens wakker met pijn op de borst. De pijn is zo erg, dat de moeder van Jan meteen de dokter laat komen.

- *Welke specifieke vragen gaat u stellen bij de anamnese?*
- *Waar gaat u op letten bij het lichamelijk onderzoek?*
- *Wat denkt u dat hier aan de hand is?*
- *Moet Jan in het ziekenhuis worden opgenomen? Waarom?*
- *Welke laboratoriumonderzoeken en specifieke onderzoekstechnieken acht u noodzakelijk? Waarom?*
- *Welke behandeling zou u voorschrijven?*
- *Waarop moet vooral gelet worden tijdens het verblijf in het ziekenhuis? Hoe lang moet Jan in het ziekenhuis blijven?*
- *Wat zijn de risico's na ontslag uit het ziekenhuis?*

De tutor heeft de antwoorden op deze vragen.

In de praktijk bleek deze aanpak zo goed te werken dat in het academiejaar 1983/1984 de resterende casus uit het gedeelte "Hart" eveneens volgens deze opzet herschreven werd. Het resultaat was dat wederom een stijgende beoordeling voor het blok. Het jaar daarna, namelijk 1984/1985, besloot de planningsgroep het blokboek ongewijzigd te laten gezien de positieve evaluatieresultaten en gezien het positieve oordeel van de planningsgroep over de gang van zaken.

5.3.2 Discussie en conclusies.

De vraag in hoeverre de onderwijsevaluatie een rol heeft gespeeld bij het veranderen van het blok dient opgesplitst te worden in twee deelvragen: "Zijn de evaluatieresultaten daadwerkelijk gebruikt?" en "Welk effect had de evaluatie op de kwaliteit van het onderwijs?"
Over onderzoek naar het gebruik van evaluatieresultaten werd in hoofdstuk 1 en in het begin van dit hoofdstuk opgemerkt dat dit onderzoek in zekere zin problematisch is. Dit type onderzoek heeft namelijk vaak een retrospectief karakter aangezien het beoogt het al dan niet optreden van veranderingen te relateren aan de uitkomsten van evaluatie-onderzoek.
Aan dit retrospectieve karakter kleeft onder andere het nadeel dat er een aanzienlijke tijdsperiode kan liggen tussen het tijdstip van onderzoek naar het feitelijk gebruik van evaluatieresultaten en de periode waarin de evaluatie plaatsvond. Dit kan tot gevolg hebben dat de evaluator en degene die de informatie gebruikt zich slechts selectief een aantal zaken herinneren die indertijd een rol speelden bij de evaluatie van bijvoorbeeld een blok. In deze casestudie hebben deze problemen ook een rol gespeeld. Desalniettemin zijn er twee concrete aanwijzingen dat de evaluatieresultaten omtrent blok 3.3 in 1981/1982 van invloed zijn geweest op de veranderingen in de opzet van het blok. De eerste aanwijzing is dat een indertijd gedane suggestie voor het herschrijven van de taken ("beschrijf een klacht of situatie, formuleer een aantal vragen en geef de antwoorden aan de tutor") ook werkelijk toegepast is. De vraag of deze suggestie bij de planningsgroep een doorslaggevende rol heeft gespeeld bij deze revisie, is uiteraard niet beantwoordbaar. Derhalve kan niet zonder meer geconcludeerd worden dat er sprake is geweest van een instrumenteel gebruik van de evaluatieresultaten. Dat er sprake was van conceptueel gebruik is echter wel aannemelijk.
Een tweede aanwijzing is dat de planningsgroep steeds het initiatief nam tot gesprekken met de projectgroep. Dit wijst erop dat de planningsgroep belang hechtte aan een goede communicatie met de projectgroep en dat de planningsgroep zich betrokken voelde bij de evaluatie van het blok. Dit zijn twee factoren die volgens Levinton & Hughes (1981) het gebruik van evaluatieresultaten positief beïnvloeden. Deze

feiten krijgen nog extra betekenis als zij bezien worden in het licht van het in hoofdstuk 1 beschreven onderzoek van Cohen (1980). Cohen liet zien dat het gebruik van evaluatie-resultaten en het effect daarvan, sterk afhankelijk is de vraag of er persoonlijk contact geweest is tussen de evaluator en de gebruiker van informatie.
Een directe causale relatie tussen evaluatie resultaat en onderwijsverbetering blijft moeilijk aantoonbaar. Wel aantoonbaar is echter dat het gebruikte evaluatie-instrument in ieder geval sensitief genoeg is om veranderingen in een blok te signaleren. Twee verschijnselen geven aanleiding tot deze conclusie. Op de eerste plaats het gegeven dat het blokboek beter beoordeeld werd nadat een aantal ingrijpende veranderingen in de opzet van het blokboek doorgevoerd waren, op de tweede plaats het feit dat de beoordeling van het blokboek identiek bleef in de periode dat het blokboek ongewijzigd gebruikt werd. Deze resultaten suggereren dat het evaluatie-instrument in principe gebruikt kan worden om veranderingen te meten.
Het lijkt dus geschikt voor gebruik in onderwijsexperimenten waarin gerichte pogingen gedaan worden onderwijsverbeteringen door te voeren en effecten daarvan in kaart te brengen.

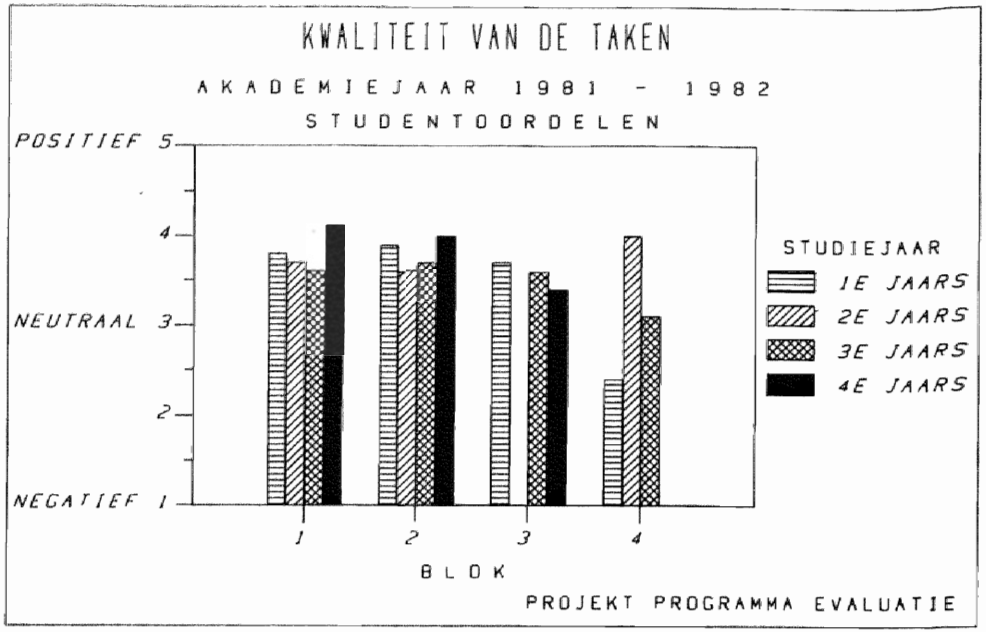
5.4 Studie 2: case blok 3.4 "leefwijzen".

In deze paragraaf wordt een tweede case-studie beschreven met betrekking tot problemen die zich in een blok voor derdejaars studenten voordeden. In deze studie wordt verslag gedaan van verschillende - ongecoördineerde - pogingen van de projectgroep en planningsgroep om de onderwijskwaliteit van blok 3.4 te verbeteren. In tegenstelling tot in de vorige studie, slaagden beide er in eerste instantie niet in om deze doelstelling te verwezenlijken. Op langere termijn trad echter een zekere verbetering op. Het uitblijven van verbeteringen in het begin zou, achteraf gezien, aan een aantal factoren toegeschreven kunnen worden: het ongunstige klimaat waarin de evaluatie uitgevoerd werd, een gebrekkige communicatie tussen planningsgroep en projectgroep, een niet doeltreffend beleid van de onderwijscommissie, en een ongunstig organisatorisch kader waarbinnen het blok gestalte moest krijgen.
De casestudie heeft betrekking op een periode van vijf opeenvolgende academiejaren, namelijk de periode 1981/1982-1985/1986. In het academiejaar 1984/1985 vond een wisseling in het coördinatorschap van de planningsgroep plaats.
Deze studie is om twee redenen van belang. In de eerste plaats toont zij de noodzaak aan van een goede communicatie tussen de evaluator en de verantwoordelijke docenten. In de tweede plaats illustreert zij het belang van een koppeling tussen onderwijsbeleid en onderwijsevaluatie.

5.4.1 Probleembeschrijving blok 3.4

Blok 3.4 (vierde blok van het derde studiejaar) staat bekend onder de naam "Leefwijzen". Het algemene thema "leefwijzen" is onderverdeeld in drie subthema's: voeding, verslaving en stress, die ieder twee weken in beslag nemen. Deze subthema's worden als zelfstandige onderdelen behandeld. In dit blok staat de wisselwerking tussen de samenleving en de gezondheid van het individu centraal. Het blok heeft een multidisciplinaire opzet. Naast biomedische disciplines als fysiologie en anatomie, klinische disciplines als klinische psychiatrie, wordt een grote inbreng geleverd door gedragswetenschappen als medische sociologie en medische psychologie.
In het begin van het academiejaar 1981/1982 besloot de onderwijscommissie van de medische faculteit om vier blokken aan te wijzen als zogenaamde "aandachtsblokken". Dit waren blokken waarvan de commissie van mening was dat zij speciale aandacht verdienden van de onderwijscommissie en de vakgroep Onderwijsontwikkeling en Onderwijsresearch. De onderwijscommissie had een aantal redenen voor dit besluit. Een voor de hand liggende reden was dat algemeen onvrede bestond over de inhoud en kwaliteit van deze blokken. Een andere reden was dat men, onder andere door jaarlijks nieuwe aandachtsblokken aan te wijzen, meer greep wilde krijgen op de kwaliteit van het onderwijsprogramma. Aandachtsblokken zouden vervolgens "diepgaand" geëvalueerd moeten worden. Een onbedoeld gevolg van deze aanpak was dat de betrokken planningsgroepen, waaronder de planningsgroep van blok 3.4, zich onder curatele gesteld voelden en dienovereenkomstig afwerend op hulppogingen reageerden. De kwalificatie "aandachtsblok" wekte aldus bij een aantal docenten wrevel en weerstand op. Dit was het klimaat waarbinnen blok 3.4 geëvalueerd werd.
In het academiejaar 1981/1982, was blok 3.4 gekwalificeerd als een van slechtst beoordeelde blokken uit het onderwijsprogramma. Het blokboek, de taken en het functioneren van de onderwijsgroep werden relatief zeer slecht beoordeeld. De evaluatieresultaten hadden op het eerste gezicht een patroon dat vergelijkbaar was met de evaluatie van blok 3.3 uit de eerste casestudie: blokboek en taken werden slecht beoordeeld en er waren problemen met het functioneren van de onderwijsgroepen. Ter illustratie is in figuur 5.3 de beoordeling van het blok, in het academiejaar 1981/1982, op de schaal "taken" weergegeven. Uit onderstaande figuur blijkt dat het blokboek van blok 3.4, zeer slecht beoordeeld werd.

Figuur 5.3 Studentoordelen takenschaal 1981/1982.



De problemen in blok 3.4 leken, zoals gezegd, in eerste instantie op de problemen zoals die in de eerste case-studie beschreven waren. Er waren echter ook een aantal duidelijke verschilpunten met blok 3.3:

- de problemen in blok 3.4 waren ernstiger dan in blok 3.3; blok 3.4 werd over het geheel genomen aanmerkelijk slechter beoordeeld dan blok 3.3.
- in blok 3.4 werd door de studenten per week aanzienlijk minder tijd aan zelfstudie besteed dan in blok 3.3: respectievelijk gemiddeld 17 en 23 uur per week.
- in blok 3.4 waren, in tegenstelling tot blok 3.3, de taken eerder te weinig gestructureerd dan teveel.
- in blok 3.4 sloot de leerstof, in tegenstelling tot blok 3.3, slecht aan op de voorkennis van studenten.
- in blok 3.4 vond een zogenaamde "Vaardigheidstoets" (zie hoofdstuk 2) plaats die een storende invloed had op de studie-aktiviteiten tijdens dit blok.

Een algemene klacht van de studenten was, dat het blok te verbrokkeld was door de behandeling van drie subthema's die ieder twee weken in beslag namen. De bezwaren van studenten richtten zich vooral op het onderwerp stress. Over het sub-

thema verslaving was men relatief gezien het meest tevreden. Uit de evaluatie bleek bovendien dat onderwijsgroepen regelmatig moeite hadden om vast te stellen wat de bedoeling van een taak was, noch duidelijk voor ogen hadden hoe een taak het beste aangepakt kon worden. Dit beeld werd bevestigd in de antwoorden van de studenten op de open vragen van de vragenlijst: "de casus stimuleren niet, werken eerder remmend", "presentatie is slecht", etcetera.
Naar aanleiding van de geconstateerde problemen werd in het evaluatieverslag de aanbeveling gedaan om, bij revisie van het blok, de aandacht te concentreren op de verbetering van de stress-taken. De overweging werd zelfs gegeven om het onderwerp stress te verplaatsen naar een ander blok, als men de klacht van studenten serieus nam dat het blok te verbrokkeld was. Bovendien deed de projectgroep het aanbod om de planningsgroep hulp te verlenen bij de hervorming van de taken.
In datzelfde academiejaar 1981/1982, kwamen om diverse redenen geen verdere contacten tot stand tussen de projectgroep en de planningsgroep. De aanbevelingen uit het evaluatierapport betreffende een nieuwe opzet voor blok 3.4 werden door de planningsgroep voor het daaropvolgende academiejaar 1982/1983 gedeeltelijk overgenomen (we laten in het midden of de beslissing voor verandering van het blok een resultaat was van de evaluatie). Het gedeelte stress werd geheel herschreven, het gedeelte over voeding bleef grotendeels ongewijzigd, en het gedeelte over verslaving werd enigzins aangepast. In tabel 5.4 zijn de veranderingen schematisch weergegeven.
Het academiejaar 1982/1983 leverde ondanks de gewijzigde opzet identieke problemen op als het voorgaande jaar. De evaluatieresultaten suggereerden zelfs een verslechtering van de situatie in het blok. In het evaluatierapport werden dezelfde aanbevelingen gedaan als het voorgaande jaar.
Op aandringen van de onderwijscommissie vond een gesprek plaats met een lid van de projectgroep om de evaluatieresultaten van blok 3.4 te bespreken. Uit dat gesprek kwam naar voren dat volgens de planningsgroep de problemen in blok 3.4 veroorzaakt werden door drie externe factoren: het plaatsvinden van de vaardigheidstoets, de gebrekkige gedragswetenschappelijke voorkennis van studenten en de geringe motivatie van studenten om psychosomatische problematiek te bestuderen. De problemen werden volgens deze visie veroorzaakt door factoren waarop de planningsgroep geen invloed kon uit- oefenen. Deze opvatting vond volgens de projectgroep echter slechts ten dele steun in de beschikbare data. Integendeel. Studenten achtten de onderwerpen die aan de orde kwam en juist wel zeer relevant en ze waren tevens van mening dat deze zaken zeker in een medische studie thuishoren (evaluatieverslagen 1981/1982 en 1982/1983).

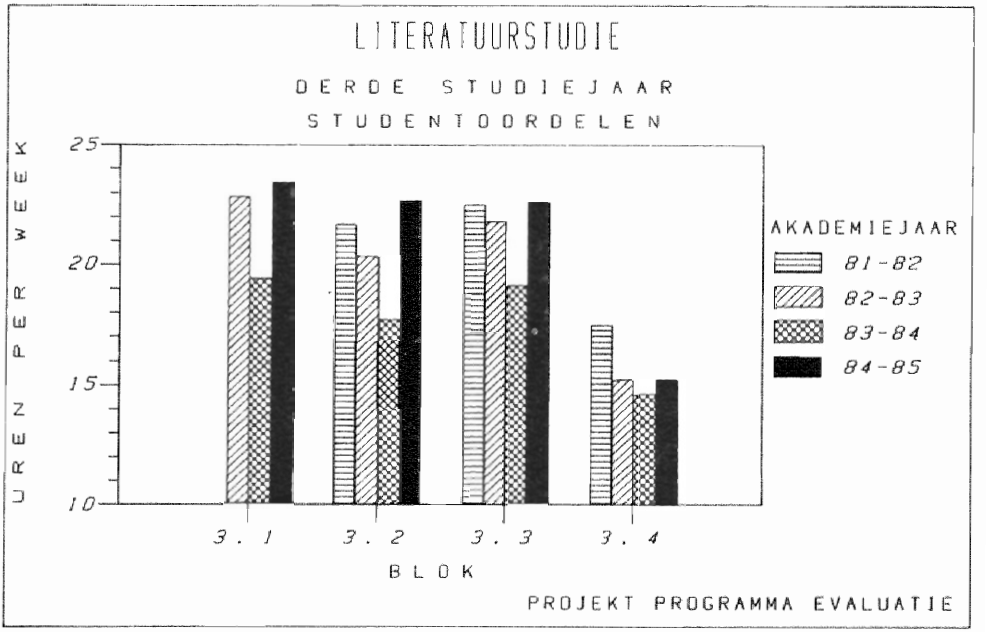
Tabel 5.4: Veranderingen in de opzet van blok 3.4 gedurende de academiejaren 1981/1982 en 1982/1983.

	1981/1982	1982/1983
Subthema: Stress	a. Inleiding op het subthema. b. 1 casus over stress bestaande uit 6 subcasus. c. 90 zelfevaluatievragen over stress. d. literatuur m.b.t. stress.	Herschreven inleiding. Nieuwe plus herziene casus. 143 zelfevaluatievragen. literatuurverwijzingen per taak.
Subthema: Voeding	a. Inleiding op het subthema. b. 3 casus over voeding. c. casus voor zelfevaluatie. d. literatuur m.b.t. voeding.	Herschreven inleiding. casus identiek aan 1981/1982 identiek aan 1981/1982 verwijzingen per taak
Subthema: Verslaving	a. Inleiding op het subthema. b. Casus over verslaving c. Zelfevaluatie. d. Literatuur m.b.t. verslaving.	identiek aan 1981/1982 Herziene casus identiek aan 1981/1982 Literatuur m.b.t. verslaving

In het gesprek tussen de evaluator en de blokcoördinator werden adviezen gegeven om de kwaliteit van het blok te verbeteren, die mede gebaseerd waren op het hierbovengenoemde rapport uit 1981/1982. Daarna vonden in dat betreffende academiejaar geen contacten meer plaats tussen planningsgroep en projectgroep.
Het jaar daarop, namelijk 1983/1984, bleek dat het blokboek van blok 3.4 aanmerkelijk beter beoordeeld werd. Uit de evaluatie kwam naar voren dat de taken in vergelijking met het voorgaande jaar meer geschikt waren voor een systematische aanpak in de onderwijsgroep, dat de taken meer aanleiding gaven tot een zinvolle groepsdiscussie en dat de taken meer aanknopingspunten gaven voor het formuleren van leerdoelen. Inspectie van het blokboek maakte duidelijk dat belangrijke veranderingen in de opzet van het blokboek waren doorgevoerd. De ingrijpendste wijziging bestond uit het schrappen van het subthema "Stress". Een andere verandering bestond uit de uitbreiding van het gedeelte dat betrekking had op voe-

ding. De bestaande casus voor zelfevaluatie werd thans gepresenteerd als casus voor de onderwijsgroep. Tenslotte werd het onderwerp "Verslaving" uitgebreid met nieuwe casus. De globaal betere studentbeoordeling bleek bij nader inzien op een aantal details een verslechtering in te houden. Dezelfde studentoordelen indiceerden namelijk dat het aantal uren zelfstudie inmiddels met drie uren gedaald was van 17.2 uren per week in 1981/1982 naar 14.2 uren per week in 1983/1984. In onderstaande figuur 5.4 is het aantal uren zelfstudie per week voor de derdejaars blokken weergegeven. De evaluatieresultaten indiceerden dat studenten, door het schrappen van het onderdeel "Stress", meer tijd kregen voor discussie in de onderwijsgroep, maar dat minder tijd aan zelfstudie besteed werd.

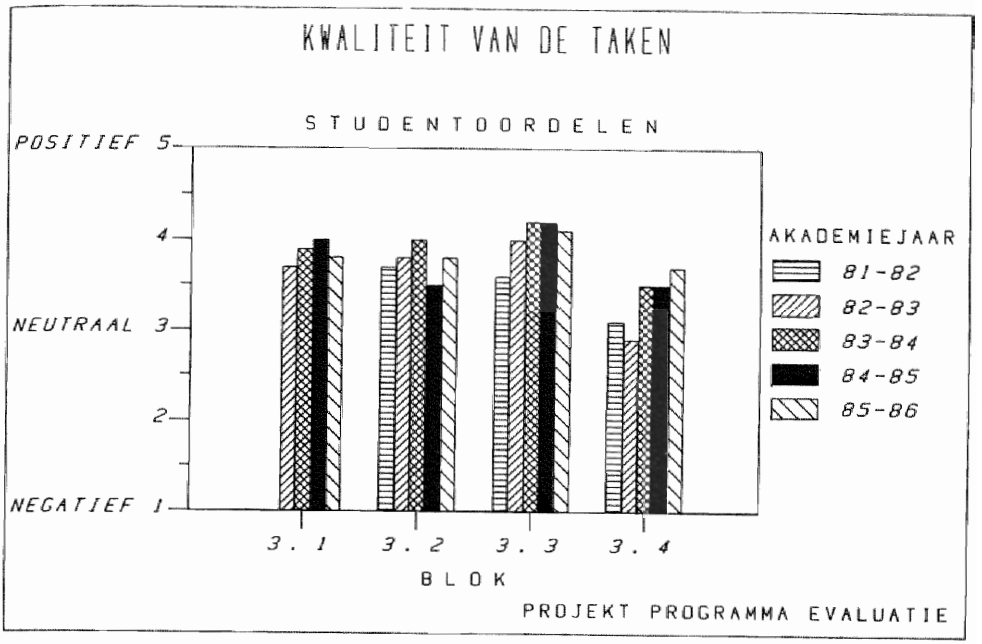
Figuur 5.4 Aantal uren zelfstudie per week, derdejaars blokken, academiejaren 1981/1982-1984/1985.



Het daaropvolgende academiejaar (1984/1985) werd een nieuwe planningsgroep geïnstalleerd die de opdracht kreeg blok 3.4 geheel te herzien en de onderdelen Verslaving, Voeding en Stress te handhaven.
De onderwijscommissie had in de tussenliggende periode besloten om de strategie rondom de zogenaamde aandachtsblokken te

verlaten. Conform de opdracht werd een nieuw blokboek geschreven door de planningsgroep.
In nauwe samenwerking tussen de projectgroep en planningsgroep werd blok 3.4 in dat jaar uitvoerig geëvalueerd.
In hetzelfde academiejaar bleek dat er wederom problemen rezen in het blok. De problemen hadden voornamelijk betrekking op de subthema's Stress en Voeding. Over het gedeelte Stress maakten de studenten de opmerking dat het onderwerp voor de derde keer in hun studie aan bod kwam en dat de taken te weinig houvast boden voor discussie in de onderwijsgroep. De taken met betrekking tot Voeding waren volgens de studenten onvoldoende gerelateerd aan het centrale blokthema. Het gedeelte Verslaving werd, evenals in vorige jaren, positief beoordeeld. In een gesprek tussen de nieuwe blokcoördinator en de evaluator werden een aantal strategieën besproken om de opbouw van het blok te veranderen. Een suggestie die werd gedaan was om het gedeelte Stress te herschrijven (daarbij rekening houdend met het feit dat het onderwerp reeds tweemaal in het programma behandeld was) en naar voren te plaatsen. Een andere suggestie was om het gedeelte Voeding zodanig aan te passen dat een betere aansluiting op het centrale thema bereikt werd.
Het daaropvolgende academiejaar 1985/1986 werd blok 3.4 wederom in nauwe samenwerking met de planningsgroep geëvalueerd. Uit de evaluatie bleek dat over het geheel genomen de waardering voor het blokboek op enkele punten iets was toegenomen (gemiddeld 0.2 stijging op de vijfpuntsschaal), maar dat voor rest geen verandering merkbaar was (het globaal oordeel over het blok bleef ongewijzigd, hetzelfde gold voor het aantal uren besteed aan zelfstudie en het oordeel over het functioneren van de onderwijsgroepen).
In onderstaande figuur 5.5 is het oordeel op de schaal "taken" weergegeven over de periode 1981/1982 - 1985/1986. In tabel 5.5 staan de bijbehorende scores en de resultaten van een F-toetsing. Uit deze figuur blijkt dat gedurende deze periode het gemiddeld oordeel is gestegen van 3.1 naar 3.5. Het studentoordeel steeg met andere woorden van neutraal naar gematigd positief. De grootste stijging deed zich voor in het academiejaar 1983/1984; het jaar waarin het onderdeel Stress geschrapt werd. Uit deze figuur wordt eveneens duidelijk dat de veranderingen in 1982/1983 (zie tabel 5.4 voor een beschrijving van de wijzigingen) geen positief effect hadden op het studentenoordeel.

Figuur 5.5 Studentoordelen takenschaal derdejaars blokken, 1981/1982-1985/1986.



Tabel 5.6: Scores op de takenschaal: gemiddelde, Standaarddeviatie, en F-toets.

Blok 3.4 Academiejaar	Gem.	SD	N	F-ratio	Df	P
1981/1982	3.0	.4	10	8.8	(4,68)	.00
1982/1983	2.9	.6	12			
1983/1984	3.5	.3	15			
1984/1985	3.5	.4	18			

5.4.2 Discussie en conclusies.

Terugkerend tot de inleiding van deze casestudie kan men concluderen dat gedurende een termijn van vijf jaren een redelijke vooruitgang geboekt werd met betrekking tot de kwaliteit van het onderwijs in blok 3.4. Na vijf jaren kwam blok 3.4 op een vergelijkbaar niveau terecht als de andere derdejaars blokken. Deze vooruitgang kwam echter niet probleemloos tot stand. Een aantal factoren leken, met name in de eerste jaren, de situatie rond blok 3.4 negatief beïnvloed te hebben.

In de eerste plaats resulteerde de beslissing van de onderwijscommissie in 1981 dat blok 3.4 aandachtsblok zou worden in een gebrekkige communicatie tussen planningsgroep en projectgroep. Dit werd versterkt door een geringe betrokkenheid van de planningsgroep bij de evaluatie (Of dit ook feitelijk werd veroorzaakt door het besluit van de onderwijscommissie laten we hier in het midden en doet voor de verdere discussie ook niet ter zake). Het resultaat was dat werd voorbijgegaan aan minstens twee condities die volgens Levinton & Hughes (1981), van belang zijn om een klimaat te creëren waarin het gebruik van evaluatieresultaten gestimuleerd wordt, namelijk: conditie 1) "relevantie" en conditie 2) "betrokkenheid van de gebruiker".
In de tweede plaats was het beleid van de onderwijscommissie in de jaren 1981/1982 en 1982/1983 niet berekend op situaties waarin uit de data van het evaluatiesysteem bleek dat zich echt ernstige problemen in een blok voordeden. Met andere woorden, in dit soort gevallen ontbrak goede afstemming van uitkomsten van evaluatie-onderzoek op maatregelen van beleidsmakers. Het gevolg was dat van de kant van de onderwijscommissie - althans achteraf gezien - geen adequate akties ondernomen werden om de problemen in blok 3.4 aan te pakken.
In de derde plaats waren de randvoorwaarden waaronder blok 3.4 gestalte moest krijgen ongunstig: een vaardigheidstoets die negatief interfereerde met de studie-aktiviteiten die van studenten verwacht werden gedurende dit blok en blokdoelstellingen die bij studenten de indruk wekken dat ze voor de derde keer in het programma met hetzelfde onderwerp geconfronteerd worden. Wat betreft de vaardigheidstoets dient opgemerkt te worden dat dit probleem zich ook in de andere studiejaren voordeden. De combinatie van een vaardigheidstoets en de vormgeving van het blok resulteerde in dit geval echter, sterker dan in andere studiejaren, in een ongunstige studiesituatie. Bovengenoemde factoren lijken het gebruik van de evaluatieresultaten negatief beïnvloed te hebben.
Uit deze casestudie kan nog een conclusie getrokken worden: namelijk het evaluatie-instrument bleek evenals in de vorige studie voldoende sensitief bleek om kwaliteit en verandering in kwaliteit van het onderwijs te signaleren.

5.5 Studie 3: Evaluatie van een experiment m.b.t. de organisatie van blokplanningsgroepen.

In deze studie wordt verslag gedaan van de evaluatie van een experiment met de organisatie van blokplanningsgroepen. Dit experiment werd met ingang van het academiejaar 1986/1987 door de onderwijscommissie van de medische faculteit in vier blokken opgestart. Doel van het experiment was om de betrokkenheid van docenten, cq tutoren, bij het onderwijs te vergroten. Dit gebeurde door tutoren mede verantwoordelijk te maken voor het samenstellen van het blokboek, het organiseren

van onderwijsactiviteiten (lezingen, workshops, practica) en het formuleren van bloktoetsitems. De verwachting was dat een grotere betrokkenheid zou leiden tot kwalitatief beter onderwijs en tot een efficiënter tijdsgebruik. Dit experiment werd geëvalueerd door de projectgroep programma-evaluatie. De evaluatiegegevens moesten de onderwijscommissie informatie verschaffen over de mogelijke continuering van het experiment.
In deze studie wordt getoond hoe het evaluatiesysteem van de medische faculteit beleidsgerichte informatie kan verschaffen aan beleidmakers, in dit geval de onderwijscommissie, teneinde beslissingen te nemen over de organisatie van het onderwijs als geheel. Deze studie is om twee redenen van belang. In de eerste plaats wordt geïllustreerd dat het evaluatiesysteem behalve aan docenten ook aan beleidmakers relevante informatie kan verschaffen. In de tweede plaats suggereren de onderzoeksresultaten wederom dat het evaluatiesysteem betrouwbare en valide informatie levert.

5.5.1 Achtergronden van het experiment.

De invulling van de rollen en functies van docenten geeft regelmatig aanleiding tot discussies binnen de medische faculteit. Een van deze discussies is uitgekristalliseerd in een notitie die in januari 1985 geschreven werd door een jaarcoördinator. Ze droeg als titel "DE ORGANISATIE VAN BLOK-PLANNINGSGROEPEN" (C.A.P. Schröer; B.O. 85-6029).
In deze notitie wordt geconstateerd dat er een aantal spanningsvelden bestaan in organisatiewijze van het onderwijs in de blokken van jaar 1 tot en met 4. Er wordt gesteld dat de gecentraliseerde werving en selectie van docenten tot een aantal problemen leidt bij de vacatureverVulling van de onderwijsrollen: planningsgroeplid, tutor, en inhoudsdeskundige. De volgende problemen worden door Schröer gesignaleerd.
"- Er is een grote spreiding in besluitvorming (planningsgroepen, capaciteitsgroepen, onderwijscommissie, deelproject training en vorming) rondom het blok, maar ook binnen het blok (planningsgroep, tutor, capaciteitsgroep/inhoudsdeskundige), waarbij de mogelijkheid tot afstemming en de verantwoordingslijnen niet altijd even duidelijk zijn.
- Het ontwerpen en uitvoeren van het onderwijs zijn in grote mate gescheiden processen waardoor de feedback aan kwaliteit verliest.
- Op individueel niveau leidt dit tot versnippering en gebrek aan samenhang tussen de diverse onderwijsactiviteiten.
- Vraag en aanbod van de bemanning zijn niet op elkaar afgestemd.

Zeer duidelijk doet dit zich voor bij benoemingen binnen planningsgroepen indien de onderwijscommissie zonder of na een afwijkend advies beslist. Een vraagprofilering ten aanzien van tutoren ontbreekt nagenoeg daar de benoeming daar-

van zich onttrekt aan de waarneming van de planningsgroep.
- De "aanvoerlijnen" van inhoudsdeskundigheid worden slechts gedeeltelijk gecontroleerd door de planningsgroep, omdat deze door haar beperkte omvang het veld onvoldoende kent, waardoor tevens veel tijd verloren gaat met het leggen, onderhouden of herstellen van kontakten.
- De betrokkenheid van inhoudsdeskundigen en vooral tutoren met de onderwijsblokken laat te wensen over, omdat zij fragmentarisch deel vormt van het geheel aan onderwijsactiviteiten. Inhoudsdeskundigen en tutoren participeren slechts gedeeltelijk respectievelijk niet in het ontwerp van het blok, zijn derhalve belemmerd in het dragen van de doelstellingen en aldus minder gemotiveerd deze te realiseren en feedback te leveren aan de planningsgroep. Men wordt in geringe mate beloond (uren-vergoeding). Tenslotte staan de inhoudsdeskundigen en tutoren niet in een duidelijke verantwoordelijkheid tot de coördinator".

Samenvattend stelt Schröer dus dat de structuur van de onderwijsorganisatie als een oorzakelijk factor gezien kan worden voor bovengenoemde problemen. Voor een helder begrip van de hier gesignaleerde problematiek is het noodzakelijk, dat een korte terugblik plaatsvindt met betrekking tot de structuur van de onderwijsorganisatie van de medische faculteit van de RL. In hoofdstuk 2 is een uitgebreide beschrijving gegeven van het onderwijssysteem en de onderwijsorganisatie van de medische faculteit.
Een van de conclusies uit dat hoofdstuk was, dat de organisatiestructuur duidelijke overeenkomsten vertoonde met het zogenaamde rationeel bureaucratisch organisatiemodel. Dit is een organisatiemodel (zie hoofdstuk 1) waarin een sterke nadruk ligt op een planmatige, rationele benadering van het onderwijs en waarin functionele scheidingen in taken van docenten gemaakt worden. Een andere conclusie in hoofdstuk 2 was, dat de beslissingsbevoegdheden van docenten over het onderwijsprogramma gelimiteerd zijn en afhankelijk van de onderwijsfuncties die zij in een blok vervullen. Tutoren hebben, bijvoorbeeld, geen enkele beslissingsbevoegdheid ten aanzien van de concrete invulling van het onderwijs in de blokken waarin zij onderwijsgroepen begeleiden. Strikt genomen vervullen tutoren slechts een uitvoerende taak. Planningsgroepen hebben daarentegen wel beslissingsbevoegdheden ten aanzien van de invulling van het onderwijs (uiteraard binnen de door de faculteit gestelde kaders), maar niet ten aanzien van de formulering van blokdoelstellingen en de keuze van tutoren. De onderwijscommissie stelt de blokdoelstellingen vast (behoudens goedkeuring van de faculteitsraad). Tutoren worden, op basis van vrije inschrijving, geworven door Buro Onderwijs. Deze strikte scheiding van beslissingsbevoegdheden binnen bepaalde onderwijsfuncties leidt ertoe dat docenten soms het gevoel krijgen dat teveel voor hun

geregeld wordt.
Genoemde notitie concentreert zich met name op de relatie tutor - planningsgroep en de bevoegdheden van docenten binnen deze onderwijsfuncties. In essentie komen de daarin gedane veranderingsvoorstellen erop neer dat tutoren voortaan dezelfde verantwoordelijkheid en bevoegdheden krijgen als planningsgroepenleden en inhoudsdeskundigen. Tutoren krijgen twee bestaande functies toebedeeld, namelijk de functie van inhoudsdeskundige en de functie van planningsgroepslid. Dit vereist derhalve dat tutoren in staat zijn om een inhoudelijke bijdrage te leveren bij de voorbereiding en de uitvoering van het blok. Voor de rest werden in deze notitie geen wezenlijke veranderingen voorgesteld in de structuur van de onderwijsorganisatie.
Met betrekking tot de uitvoering van de voorgestelde veranderingen werden de volgende maatregelen genomen. Iedere tutor moest twee onderwijsgroepen begeleiden. Een blok bevat normaliter 18 onderwijsgroepen, hetgeen betekent dat planningsgroepen aldus zouden bestaan uit een coördinator, een of twee student-leden en negen tutoren (annex planningsgroepsleden). De coördinatie en uitwerking van het blokboek werd in handen gelegd van de coördinator en een adjunct-coördinator. De overige planningsgroepleden werden betrokken bij de constructie van taken, de formulering van toetsitems, de selectie van leermiddelen, etcetera. De volgende voordelen werden van het experiment verwacht:
- een verkorting tot een minimum van de lijnen tussen planningsgroep, tutoren en inhoudsdeskundigen;
- een verminderde versnippering van krachten;
- een efficientere functievervulling;
- grotere betrokkenheid met en motivatie voor zowel ontwerp als uitvoering bij docenten;
- toename en verbetering van feed-back.

Om voor de faculteit de zaken werkbaar te houden, wordt voorgesteld de nieuwe organisatievorm als experiment in een beperkt aantal blokken toe te passen. Een begeleidingscommissie, bestaande uit vertegenwoordigers van de onderwijscommissie, onderwijskundigen en jaarcoördinatoren, moet zorgdragen voor de voorbereiding, uitvoering en evaluatie van het experiment.

5.5.2 Procedure.

Het experiment werd in het academiejaar 1986/1987 in vier blokken uitgevoerd: blok 3.1 (infekties en ontstekingen), blok 1.2 (traumata), blok 4.3 (gynaecologische problemen en zwangerschapsproblemen), en blok 2.4 (de adolescent). Per studiejaar werd een blok gekozen. Kenmerkend voor de betreffende blokken was, dat het volgens de onderwijsevaluatie redelijk tot goed functionerende blokken waren en dat geen

aanleiding bestond tot grootscheepse veranderingen. Studenten wisten dat in deze blokken geëxperimenteerd werd met de blokorganisatie.
De evaluatie van het experiment was op twee zaken gericht, namelijk op het meten van veranderingen in de kwaliteit van het onderwijs en op het onderzoeken van het functioneren van de nieuwe planningsgroepen.
Ten aanzien van het eerste was de aandacht vooral gericht op het functioneren van de tutoren, onderwijsgroepen, en de kwaliteit van het blokboek. Ten aanzien van het tweede was de aandacht gericht op de hierboven genoemde vooronderstelde voordelen, c.q. beoogde effecten van de experimentele organisatievorm.
De kwaliteit van het onderwijs werd op de gebruikelijke manier, met behulp van de standaardvragenlijst voor studenten, gemeten. Het functioneren van planningsgroepen werd geëvalueerd met behulp van een speciaal voor dat doel ontwikkelde vragenlijst. Deze lijst werd na afloop van het blok aan de planningsgroepsleden c.q. tutoren voorgelegd. De vragenlijst bevatte items over de achtergrond van de tutoren (ervaringen met eerdere tutorschappen, en planningsgroepen), het functioneren van de planningsgroep, het zelfoordeel over de begeleiding van de onderwijsgroepen, etcetera. In bijlage 4 is een voorbeeld van de betreffende vragenlijst opgenomen.
Een voor de hand liggende vraag was of door de experimentele organisatievorm reële verbeteringen in de kwaliteit van het onderwijs optraden. Daartoe werden van de experimentele blokken de gemiddelde schaalscores van de standaardvragenlijst vergeleken met de gemiddelde schaalscores van dezelfde blokken uit het voorgaande academiejaar. Deze blokken dienden dus als controle op het effect van de interventie. Hetzelfde gebeurde op itemniveau. Met behulp van variantie-analyse werd statistisch getoetst of er significante verschillen bestonden in de beoordeling van de experimentele blokken en de controle blokken. De statistische toetsing van de verschillen tussen de gemiddelde schaalscores vond plaats met behulp van een F-toets. De eenheid van analyse was de gemiddelde score van een onderwijsgroep. In ieder blok hadden 18 onderwijsgroepen plaats. Met behulp van éénweg-variantie-analyse werden per blok F-ratio's berekend. Getoetst werd de hypothese H_0, dat er geen verschillen waren tussen de gemiddelde scores van een experimenteel blok en een controleblok.

5.5.3 Resultaten en discussie.

In tabel 5.7 zijn de resultaten uit de toetsing van de schaalscores samengevat. Per blok zijn alleen dan de gemiddelde scores en standaarddeviatie weergegeven als er statistisch significante verschillen ($p < 0,05$) waren tussen het experimentele blok en het controleblok. Uit de hoeveelheid lege cellen in tabel 5.7 blijkt, dat het merendeel van de

schaalscores niet van elkaar verschilde. Slechts op enkele schalen werden significante verschillen gevonden.

Tabel 5.7: Samenvatting resultaten onderwijsevaluatie.

Blok	3.1				1.2				4.3				2.4			
Academiejaar	85/86		86/87		85/86		86/87		85/86		86/87		85/86		86/87	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
1. Taken																
2. Tutor	3.9	.4	4.2	.3	3.5	.7	4.0	.5								
3. Onderwijs-groep																
4. Skillslab					4.2	.2	4.0	.2	3.9	.2	4.3	.2				
5. Zwaarte																
6. Bloktoets	3.5	.2	3.7	.2												
7. Leermiddelen									3.6	.3	3.2	.3	3.8	.3	4.0	.2
8. Globaal Oordeel																
10. Inhoudsdes.																
11. Sociale Vaardigheden																
12. Onafh. Studie																
13. Structurering																
14. Afwisseling					3.9	.2	4.2	.3	3.9	.3	4.2	.3				
N (aantal onderwijsgroepen)	18		18		18		18		18		18		18		18	

Uit tabel 5.7 blijkt dat in twee blokken (3.1 en 1.2) significante verschillen optraden in de beoordeling van tutoren (schaal 2). De gemiddelde beoordeling op de tutorschaal stijgt in blok 3.1 met 0,3 schaalwaarde en in blok 1.2 met 0,5 schaalwaarde. Inspectie van de itemscores van deze schaal laat zien dat op het merendeel van deze items sprake is van een gemiddelde stijging van respectievelijk 0,3 en 0,5. De tutoren worden in de experimentele blokken 3.1 en 1.2, in vergelijking met het jaar daarvoor, aanmerkelijk positiever beoordeeld.
In de blokken 4.3 en 2.4 was geen sprake van statistisch significante verschillen op de tutorschaal. In blok 4.3 werd een gemiddelde stijging van 0,2 gevonden. In blok 2.4 bleken de tutoren in het experimentele blok 0.2 lager beoordeeld te zijn. Verdere significante veranderingen in de beoordeling van de experimentele blokken worden gevonden in de skillslab-schaal (medisch-technisch), de leermiddelenschaal, en de afwisselingschaal. De dalende beoordeling voor programma-onderdelen van het skillslab kan niet verklaard worden door het experiment, aangezien planningsgroepen geen invloed op dat programma hebben.
Bovenstaande resultaten lijken wat betreft het functioneren

van tutoren één conclusie in ieder geval te rechtvaardigen: het experiment is in twee blokken in zijn opzet geslaagd. Door het vergroten van de verantwoordelijkheid bij tutoren en aldus een grotere betrokkenheid te bewerkstelligen werden de tutoren in deze blokken beter beoordeeld. Blok 4.3 lijkt een twijfelgeval te zijn. De scores tenderen naar een positievere beoordeling maar zijn niet significant. In blok 2.4 lijkt het experiment op het eerste gezicht mislukt; althans wat betreft de verwachting dat tutoren positiever beoordeeld zouden worden. Uit de gemiddelde scores blijkt echter dat deze in vergelijking met de andere experimentele blokken positief zijn. Men zou bovenstaande conclusie kunnen aanvechten door te wijzen op het Hawthorne-effect: studenten en tutoren zouden beïnvloed kunnen zijn door de experimentele situatie sec met als gevolg dat studenten de kwaliteit van het onderwijs (tutor, onderwijsgroep, blokboek, etcetera) in het betreffende blok positiever waarnamen dan door de feiten gerechtvaardigd werd. Uit de gegevens in tabel 5.7 blijkt echter dat deze redenering niet steekhoudend is. In een dergelijk geval zouden immers meer positieve verschillen te verwachten zijn dan thans het geval is.
De verwachting dat onderwijsgroepen beter gaan functioneren en dat blokboeken beter beoordeeld worden, komt niet uit. Het is op zich merkwaardig dat in blokken waarin tutoren significant beter beoordeeld worden er geen sprake is van een significante stijging in de beoordeling van onderwijsgroepen. Misschien dat de correlatie tussen het functioneren van tutoren en onderwijsgroepen te laag was om merkbare effecten op de onderwijsgroep te meten. Het is niet uitgesloten dat dat het geval was. Uit onderzoek is namelijk gebleken dat de gemiddelde correlatie tussen het functioneren van tutoren en het functioneren van onderwijsgroepen .32 bedraagt (Gijselaers & Schmidt, 1985b). De beoordeling van de blokboeken neemt eveneens niet toe. In twee blokken is wel sprake van een toegenomen variatie in taakvormen en behandelde onderwerpen. De hypothese dringt zich op dat grotere planningsgroepen, met meer inbreng uit verschillende disciplines, in sommige blokken wel meer variatie in taakvormen en onderwerpen kunnen realiseren maar dat geen meetbare effecten optreden in de kwaliteit van de taken.
Een voor de hand liggende vraag is of, zoals in de notitie werd gesuggereerd, de experimentele planningsgroepen efficiënter en met meer betrokkenheid vorm hebben gegeven aan het onderwijs in de blokken. De vraag of het resultaat daarvan betere blokboeken opleverde, is hiervóór reeds beantwoord. Het functioneren van de experimentele planningsgroepen werd geëvalueerd met een voor dat doel ontwikkelde vragenlijst voor tutoren. Tabel 5.8 bevat in rechte tellingen de zelfoordelen van tutoren/planningsgroepsleden op een aantal vragen over hun betrokkenheid bij het blok. Tabel 5.9 bevat de antwoorden op een open vraag aan de tutoren betreffende hun

positieve en negatieve ervaringen met de experimentele blokorganisatie te beschrijven. Uit de zelfoordelen in tabel 5.8 blijkt dat de tutoren over het geheel genomen positief zijn over het experiment. De belangrijkste conclusies met betrekking tot de oordelen van tutoren zijn:

- de tutoren zijn van mening dat op deze manier efficiënt gewerkt kan worden;
- de tutoren voelden zich meer dan in andere blokken, betrokken bij het blok;
- de tutoren waren over het geheel genomen beter in staat om de onderwijsgroep in de richting van de doelstellingen te leiden;
- de tutoren waren over het geheel genomen in staat om gerichte vragen in de onderwijsgroep te stellen;
- de tutoren hadden over het geheel genomen meer plezier in hun rol;
- dertig van de vierendertig tutoren is van mening dat de FdG meer blokken op deze wijze moet organiseren.

Tabel 5.8: Oordeel van planningsgroepsleden m.b.t. de nieuwe organisatievorm.

		Volledig oneens				Volledig Eens	
		1	2	3	4	5	
17.	Ik vond deze manier van werken efficiënt	1	2	-	18	19	N=37
18.	In vergelijking met andere blokken voelde ik me in dit blok als tutor meer betrokken	3	-	2	12	15	N=30
19.	Ik kon de onderwijsgroep in dit blok, in vergelijking met vorige blokken, beter leiden in de richting van de doelstellingen van het blok	2	5	5	11	9	N=32
20.	Ik was in vergelijking met vorige blokken, beter in staat om gerichte vragen te stellen in de onderwijsgroep	2	5	9	9	8	N=33
21.	Ik had als tutor, in vergelijking met vorige blokken meer plezier in mijn rol.	2	7	6	9	9	N=33
22.	Zou U de FdG adviseren meer blokken op deze wijze te organiseren.	Nee 1		Ja 30		Weet niet 3	N=34

In tabel 5.9 zijn de positieve en negatieve ervaringen van de planningsgroepsleden weergegeven. Uit de antwoorden blijkt dat men relatief meer positieve dan negatieve ervaringen heeft met de experimentele blokorganisatie, dat men meer betrokkenheid bij het blok voelt, dat de tutorrol beter uitvoerbaar is en dat men directe terugkoppeling krijgt over het functioneren van taken. De negatieve ervaringen hebben betrekking op de belasting van het begeleiden van een dubbele onderwijsgroep (twee groepen achter elkaar), de ondoelmatigheid van vergaderingen met volledige planningsgroepen en met tutorinstructies die niet toereikend zijn voor niet-medisch geschoolden.

Tabel 5.9: Ervaringen van planningsgroepsleden met het organisatiemodel, op basis van inventarisatie in 4 experimentele blokken.

POSITIEF ERVAREN.

- efficiënter werken (2x)
- meer betrokkenheid tutor (9x)
- tutorrol beter uitvoerbaar (5x)
- directe terugkoppeling over het functioneren van de taken (5x)
- als tutor verantwoordelijk voor volgende versie (1x)
- verhoogde opmerkzaamheid bij probleemsituatie (1x)

NEGATIEF ERVAREN.

- volledige planningsgroep niet altijd efficiënt (3x)
- voorbereidingstijd te kort (1x)
- dubbele onderwijsgroep te grote belasting (3x)
- tutorinstructie houdt geen rekening met niet-medisch geschoolden (3x)
- geen bevredigend overleg over bloktoets (1x)
- coördinator is grotere sleutelfiguur (1x)

5.5.4 Conclusies.

Naar aanleiding van deze studie lijken twee conclusies gerechtvaardigd.
In de eerste plaats geven de verzamelde data geen aanleiding om het experiment te beëindigen. Noch in de onderwijsevaluatie noch in de speciale vragenlijst voor tutoren worden gegevens gevonden die erop wijzen dat het experiment negatieve effecten heeft op de betrokkenheid van docenten of op de kwaliteit van het onderwijs. De data geven de indruk dat docenten met meer plezier en betrokkenheid meegewerkt hebben

aan de voorbereiding en uitvoering van het onderwijs. Een indicatie voor deze conclusie wordt onder andere gegeven in de antwoorden op de vraag aan docenten of zij de onderwijscommissie zouden adviseren meer blokken op deze wijze te organiseren. Het antwoord was met een overgrote meerderheid "ja". De onderwijscommissie heeft dit advies gedeeltelijk overgenomen door te besluiten om het experiment te continueren. Men kan bovengenoemde conclusie wellicht zelfs omdraaien: de data geven aanleiding om het experiment te continueren en uit te breiden. Er lijken aanwijzigingen te zijn dat de experimentele organisatiewijze positieve effecten heeft op het functioneren van tutoren, in enkele gevallen op de kwaliteit van de bloktoets en de afwisseling in taakvormen en onderwerpen in de blokboeken.
In de tweede plaats blijkt dat de evaluatie-instrumenten bruikbare informatie geven over veranderingen in de kwaliteit van het onderwijs.

5.6 Studie 4: Een verkennend onderzoek naar de stabiliteit van tutorgedrag.

5.6.1 Inleiding.

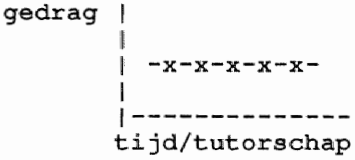
De nu volgende studie heeft, evenals studie 3, betrekking op het functioneren van tutoren. Vormden praktische vragen van docenten of beleidsmakers de directe aanleiding tot de vorige studies, deze studie heeft een meer fundamenteel karakter.
Uit studie 3 bleek dat een gewijzigde organisatie van het blokonderwijs de betrokkenheid van tutoren vergrootte. In de nu volgende studie wordt het normale dagelijkse functioneren van tutoren onderzocht.
Zoals bleek uit studie 3, hecht de medische faculteit belang aan het goed functioneren van tutoren. De tutor wordt als medebepalend gezien voor de kwaliteit van het onderwijs.
De faculteit heeft een aantal maatregelen genomen om de kwaliteit van het functioneren te reguleren. In de eerste plaats zijn docenten, voordat zij bijdragen gaan leveren aan het onderwijs, verplicht om twee cursussen te volgen (een inleidende cursus over het onderwijssysteem als geheel en een cursus specifiek gericht op de taken van tutoren), alvorens zij onderwijsgroepen mogen begeleiden. In de tweede plaats ontvangen tutoren van de projectgroep programma-evaluatie, na afloop van een blok steeds een evaluatierapport waarin feedback wordt gegeven over hun functioneren.
Uit onderzoek is echter gebleken dat eenmalige cursussen waarin docenten bepaalde technieken geleerd krijgen, nauwelijks effect hebben op de kwaliteit van het onderwijs (Levinson-Rose & Menges, 1981). Bovendien is bekend dat feedback tot geen aantoonbare verbeteringen van de onderwijskwaliteit leidt als deze beperkt blijft tot schriftelijke rapportage (Cohen, 1980). In de medische faculteit heeft men o.a. getracht deze problemen te ondervangen door de ontwikkeling van herhaaltrainingen te stimuleren (Moust, de Grave & Gijselaers, 1985). Men heeft bijvoorbeeld overwogen om beginnende tutoren na twee of drie tutorschappen een verplichte herhaalcursus te laten volgen. Andere maatregelen die overwogen zouden kunnen worden, bestaan hieruit dat slecht functionerende tutoren eveneens een dergelijke bijscholing moeten volgen, of dat deze tutoren op andere wijze in het onderwijs worden ingezet.
Een laatste maatregel zou kunnen zijn dat feedback, bij minder goed functionerende of beginnende tutoren, altijd gepaard gaat met een gesprek tussen evaluator en tutor. Tot de uitvoering van dergelijke maatregelen is het echter nooit gekomen. Factoren als tijdgebrek, gebrek aan menskracht en andere prioriteiten in het onderwijsbeleid van de faculteit speelden een rol bij het uitblijven daarvan. Daarbij komt nog dat dergelijke maatregelen pas zin hebben als informatie beschikbaar is over het feitelijke optreden van tutoren in het onderwijs. Is het zo dat er inderdaad zwakke tutoren bestaan? Leren die tutoren van hun ervaring? Met andere

woorden: Gaan ze het beter doen naarmate ze vaker ingezet zijn? In hoeverre is het optreden van tutoren stabiel over verschillende situaties? Inzicht in deze problematiek is noodzakelijk voordat men adequate beleidsmaatregelen in deze kan nemen.

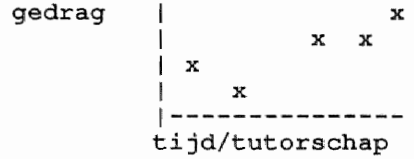
5.6.2 Enkele methodologische overwegingen bij onderzoek naar de stabiliteit van docentengedrag.

In deze paragraaf wordt een beknopt overzicht gegeven van methodologische aspecten die aan onderzoek naar de stabiliteit van docentengedrag verbonden zijn.
Rogosa, Floden en Willet (1984) onderscheiden twee onderzoeksvragen over de stabiliteit van docentengedrag in een bepaalde tijdsperiode:
Vraag 1: Is het gedrag van individuele docenten consistent binnen een bepaalde tijdsperiode?
Vraag 2: Zijn individuele verschillen tussen docenten consistent binnen een bepaalde tijdsperiode?
De eerste vraag kan men vertalen als: leren tutoren van hun eerdere ervaringen of blijft hun optreden constant in de loop van de tijd?

figuur 5.6

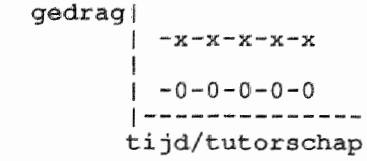


figuur 5.7

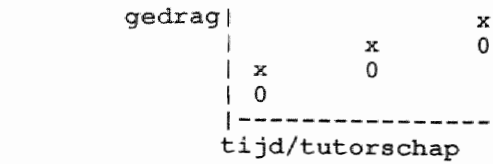


In figuur 5.6 en 5.7 zijn als voorbeeld de beoordelingen van een tutor als functie van de tijd weergegeven. In figuur 5.6 is sprake van perfecte temporele stabiliteit. De betreffende tutor krijgt steeds dezelfde beoordeling. In figuur 5.7 is sprake van een opwaartse trend in beoordeling als tutor met soms afwijkingen ten opzichte van die trend (scatter).
De twee vraag heeft betrekking op de rangorde van tutoren ten opzichte van elkaar: bijvoorbeeld de vraag of een relatief goed beoordeelde tutor binnen een bepaalde periode relatief goed beoordeeld blijft. Imaginaire data die een illustratie vormen van deze vraag, zijn in figuur 5.8 en 5.9 weergegeven. Wederom zijn tutorbeoordelingen als functie van de tijd weergegeven. In beide figuren is sprake van perfect consistente verschillen tussen een twee tutoren, binnen een aantal tutorschappen, zij het dan dat in figuur 5.8 tutoren niet lijken te leren van eerdere ervaringen, terwijl zij dat in figuur 5.9 wel lijken te doen.

figuur 5.8



figuur 5.9



5.6.3 Statistische procedures voor de analyse van vraag 1.

Guire en Kowalski (1979) beschrijven een aantal statistische procedures waarmee gedrag als functie van de tijd geanalyseerd kan worden. Gegeven een reeks metingen O_1, O_2, O_3, O_4, ... O_t van een individu, wordt verondersteld dat een meting op tijdstip T in de volgende vorm kan worden weergegeven:

$$O_t = f(t) + e_t.$$

Volgens deze vergelijking is een meting O_t op tijdstip T, afhankelijk van een systematische component f_t, en een stochastische component e_t. Het probleem is om een functie f_t te vinden die een optimale fit heeft met de metingen O_t. Een voorbeeld van een dergelijke functie f_t is hieronder weergegeven:

$$f(t) = \beta\, O_{t-1}.$$

Volgens deze functie, die men een Markov-functie noemt, is gedrag op tijdstip t afhankelijk van gedrag op tijdstip t-1. Alvorens men gaat zoeken naar een dergelijke functie f_t is het noodzakelijk dat men stochasticiteit in de tijdreeks kan uitsluiten (Frederiksen & Rotondo, 1979). Twee eenvoudige methoden om een tijdreeks op stochasticiteit te toetsen zijn: enkelvoudige regressie-analyse en nonparametrische correlatie-analyse. Gegeven de observaties O_1, O_2, ...O_t op de tijdstippen 1, 2.....t kan men de volgende getallenparen vormen: $(1,O_1)$, $(2,O_2)$,.....(t,O_t).
Deze getallenparen kunnen in een assenstelsel weergegeven worden met op de x-as de variabele "tijd" en op de y-as de afhankelijke variabele.
Bij een simpele regressie-analyse fungeert de variabele tijd als onafhankelijke variabele en de variabele gedrag als afhankelijke. De product-moment correlatie r_{xt} is een indicator voor de associatie tussen een lineaire tijdtrend en gedrag. Als men de beschikking heeft over veel observaties kunnen nonlineaire regressie-analyses verricht worden. Als de tijdreeks een stochastisch patroon vormt, is de verwachte waarde van de product-moment correlatie R of de rangordecorrelatie (Spearman's rho of Kendall's tau) tussen O_t en T,

gelijk aan nul. Als de reeks een lineaire trend vertoont is de product-moment correlatie tussen O_t en T significant afwijkend van nul. Toetsing van de nulhypothese R = 0, rho = 0 of tau = 0, geeft antwoord op de vraag of de reeks een opwaartse of neerwaartse trend vertoont. Als de nulhypothese verworpen wordt, kan men op zoek gaan naar een lineaire functie f_t die een beschrijving geeft van de metingen O_t. In deze studie werd de product-moment correlatie R tussen tutor-functioneren en tijdstip van beoordeling berekend.

5.6.4 Statistische procedures voor de analyse van vraag 2.

Rogosa, Floden en Willet (1984) noemen drie technieken om de consistentie van individuele verschillen binnen een bepaalde tijdsperiode te bepalen:

a. berekening van de correlatie tussen observaties op opeenvolgende tijdstippen;
b. toepassing van de generaliseerbaarheidstheorie;
c. toepassen van herhaalde metingen variantie-analyse designs.

Deze statistische technieken zijn regelrecht afgeleid uit methoden om de betrouwbaarheid van meetinstrumenten te bepalen (zie hoofdstuk 3 en hoofdstuk 4). De achterliggende gedachtengang is dat metingen binnen een tijdreeks op te vatten zijn als herhaalde metingen van individuen.
Naarmate de herhaalde metingen meer consistente resultaten opleveren, is er sprake van toenemende consistentie van individuele verschillen tussen docenten.
In deze studie werd, vanwege zijn eenvoud, de eerste techniek gebruikt. Consistenties in verschillen tussen individuen worden volgens deze methode bepaald door gebruik te maken van correlaties tussen observaties op opeenvolgende tijdstippen. Dergelijke correlaties zijn vergelijkbaar met test-hertest betrouwbaarheden.
Beoordelingen op tijdstip 1 worden gecorreleerd met die op tijdstip 2. Naarmate de correlatie tussen beoordelingen op twee tijdstippen stijgt, neemt de test-hertest betrouwbaarheid toe: de rangorde in de scores op tijdstip 1 lijkt meer op de rangorde op tijdstip 2 naarmate de correlatiecoëfficiënt de waarde 1 benadert. Als men de stabiliteit over meerdere tijdstippen wil bepalen, wordt de mediane of de gemiddelde correlatie van alle correlaties tussen de afzonderlijke tijdstippen berekend. Voorbeelden van onderzoek waarin deze techniek gebruikt werd, zijn beschreven door Shavelson en Dempsey-Atwood (1976). Zij vermelden dat in de meeste onderzoeken correlaties tussen docentengedrag op tijdstip 1 en tijdstip 2 gevonden worden met een waarde van .17. Onderzoek waarin de stabiliteit over meerdere tijdstippen bepaald wordt, leverde mediane correlaties rondom nul op.

5.6.5 Procedure.

Het onderzoek had betrekking op 326 tutoren die ieder, in de academiejaren 1981/1982 - 1984/1985, meer dan één tutorschap vervuld hadden. In totaal hadden de 326 tutoren 1002 tutorschappen vervuld. Alle tutoren ontvingen na afloop van een blok een evaluatierapport waarin feedback gegeven werd over hun functioneren als begeleider van hun onderwijsgroep. De projectgroep programma-evaluatie beoogt uiteraard met dergelijke rapporten tutoren feedback te geven in de veronderstelling dat, indien studenten een in verhouding negatief of neutraal oordeel over de tutor geven, tutoren deze informatie gebruiken om in hun daaropvolgende tutoroptreden hun gedrag te wijzigen in de gewenste richting. Dit rapport, de zogenaamde "tutorbeoordeling", bevat daarom, naast een beoordeling van alle studenten die deel uit maakten van de betreffende onderwijsgroep, een gemiddelde beoordeling van alle tutoren in het betreffende blok, wat de tutoren de mogelijkheid biedt hun persoonlijke beoordeling te toetsen aan de gemiddelde beoordeling van andere tutoren. In bijlage 5 is een voorbeeld van een (gefingeerde) tutorbeoordeling opgenomen. Konsekwenties zijn, zoals gezegd, niet aan de beoordeling verbonden. Feedback aan tutoren blijft beperkt tot een schriftelijke rapportage; normaliter vinden geen contacten plaats tussen evaluator en docent om de beoordeling te bespreken.

5.6.6 Onderzoeksopzet.

In het onderzoek werden de tutoren verdeeld over tien subgroepen, afhankelijk van het aantal onderwijsgroepen dat die tutoren in de onderzoeksperiode begeleid hadden. Iedere subgroep vertegenwoordigde groepen tutoren die respectievelijk 1, 2, 3, 4, 5, 6, 7, 8, 9, of 10 onderwijsgroep(en) begeleid hadden.
Een tijdreeksdesign werd opgebouwd dat bestond uit tien groepen. De eerste groep tutoren (N= 109) had één tutorschap afgerond, de tweede groep twee, etcetera. Groep I fungeerde als controlegroep. In onderstaande figuur is het design weergegeven. De letter O staat voor "onderwijsgroep", de letter x staat voor de feedback die de tutoren na afloop van een blok ontvingen. De variabele "tijd" is uitgedrukt in het aantal tutorschappen dat door de tutoren werd ingevuld. De tijdsvariabele kan in dit design ook opgevat worden als een ervaringsvariabele (de hoeveelheid ervaring van docenten als tutor, uitgedrukt in het aantal vervulde tutorschappen). Dit design is derhalve geen tijdreeks-design in de meest stricte betekenis, te weten: een reeks voormetingen op vaste tijdstippen (met gelijke tijdsintervallen), een interventie gedurende een bepaalde periode, en een reeks nametingen op vaste tijdstippen (eveneens met gelijke tijdsintervallen).

figuur 5.10 Groepsindeling naar aantal tutorschappen per tutor.

Tutorschap	1	2	3	4	5	6	7	8	9	10	
groep I	O	x									N= 109
groep II	O	x	O	x							N= 56
groep III	O	x	O	x	O	x					N= 50
groep IV	O	x	O	x	O	x	O	x			N= 30
groep V	O	x	O	x	O	x	O	x	O	x	N= 32
groep VI	O	x	O	x	O	x	O	x	O	x	O x N= 20
groep VII	O	x	O	x	O	x	O	x	O	x	O x N= 15
groep VIII	O	x	O	x	O	x	O	x	O	x	O x N= 6
groep IX	O	x	O	x	O	x	O	x	O	x	O x O x N= 2
groep X	O	x	O	x	O	x	O	x	O	x	O x O x O x O x N= 6

Men kan dit design ook opvatten als een reeks metingen waarbij het aantal individuen afneemt naarmate er meer metingen hebben plaatsgevonden. Het design krijgt dan de volgende vorm:

Figuur 5.11: Tijdreeksdesign zonder indeling in subgroepen.

Tutorschap	1	2	3	4	5	6	7	8	9	10
	O x	O x	O x	O x	O x	O x	O x	O x	O x	O x

Het eerste design is met name geschikt voor longitudinale analyse van gegevens binnen subgroepen; bijvoorbeeld ter vergelijking van curves tussen subgroepen, of bij onderzoek naar de vraag of curves een bepaalde trend vertonen per subgroep. Het tweede design leent zich voor het maken van vergelijkingen tussen tijdstippen; bijvoorbeeld bij onderzoek naar de vraag of beoordelingen op latere tijdstippen hoger zijn dan op eerdere tijdstippen. Het tweede design veronderstelt dat er geen verschillen tussen subgroepen bestaan op één bepaald tijdstip, met andere woorden: het veronderstelt dat de scores op een bepaald tijdstip onhankelijk zijn van het aantal tutorschappen dat docenten vervuld hebben. Alvorens nadere analyses met behulp van dit tweede design te verrichten werd deze assumptie getoetst. Daartoe werd per tijdstip een éénweg-variantie-analyse verricht met als onafhankelijke variabele de subgroepindeling en als afhankelijke variabele de score op een item van de standaardvragenlijst dat betrekking had op het functioneren van de tutor. De nulhypothese was dat er geen verschillen waren op de afzonderlijke items tussen groepen binnen een tijdstip. Uit de variantie-analyse bleek dat er geen significante verschillen tussen de groepen bestonden zie bijlage 6. Met andere woorden de subgroepen waren onderling niet verschillend m.b.t. de tutor-items binnen een tijdstip. Dat betekent dat het tweede design gebruikt kan worden voor analyses waarin scores tussen tijdstippen met elkaar vergeleken kunnen worden.

5.6.7 Resultaten en discussie.

Onderzoeksvraag 1: Is het functioneren van tutoren stabiel over de tijd?

De eerste onderzoeksvraag betrof de vraag of het functioneren van tutoren stabiel is over een bepaalde tijdsperiode. De vraag naar het optreden van eventuele relaties tussen ervaring van tutoren en studentoordelen over hun functioneren werd onder andere benaderd volgens de hierboven beschreven statistische procedure waarin per item van de vragenlijst met betrekking tot het functioneren van de tutor de correlaties berekend worden tussen de tijds- of ervaringsvariabele en elk item van de vragenlijst afzonderlijk. Deze procedure werd toegepast binnen de afzonderlijke subgroepen en op de totale groep tutoren. Een andere benadering was dat, met behulp van het tweede design, de scores tussen tijdstippen met elkaar vergeleken werden. In onderstaande tabel 5.10 zijn per subgroep en over het totaal, de correlaties tussen studentoordelen en de tijdsvariabele weergegeven.
De itemnummers verwijzen naar de in bijlage 1 opgenomen vragenlijst voor studenten.

Tabel 5.10: Pearson's product-moment correlaties, per subgroep en over het totaal, tussen studentoordelen en de tijdsvariabele.

	Subgroep									
	2	3	4	5	6	7	8	9	10	Totaal
Vraag										
v33	09	04	15	-06	-01	10	-12	-11	-10	03
v34	-03	08	13	03	-05	00	07	-15	-28	06
v35	03	05	09	02	05	-02	-17	-31	-33"	-05
v36	04	-04	-06	-04	01	16	02	05	-25	-02
v37	07	04	11	00	-03	-01	-07	12	-24	-02
v38	05	-04	01	-06	-17	09	-20	-08	-19	-12"
v39	-02	-07	02	00	-01	05	-03	08	00	01
v40	00	-12	01	-06	02	-02	05	11	-10	02
v41	-09	-05	01	-08	02	-10	-25	-39	-28	-04
v42	03	-05	05	00	17	-09	-03	-12	-06	-01
v43	01	00	00	-01	01	-13	-02	-18	-25	05
v44	06	04	11	04	01	03	-05	-27	-21	-01
N	109	145	115	157	116	102	47	17	59	975

Noot. Correlaties met - " teken - zijn statistisch significant ($p < .01$).

Uit bovenstaande tabel blijkt dat alle correlaties, op twee na, niet significant verschillen van nul. Deze resultaten zijn een aanwijzing voor de stochasticiteit van de tijdsreeksen. Er is geen samenhang tussen de hoeveelheid ervaring van tutoren, uitgedrukt in het aantal tutorschappen, en de beoor-

deling van hun gedrag door studenten. Met andere woorden, naarmate de ervaring van tutoren toeneemt is er geen sprake van een systematische stijging of daling in studentoordelen op de items betreffende het functioneren van tutoren.
Een aantal verklaringen zijn mogelijk voor dit resultaat. De eerste is dat er sprake is van "ceiling-effecten": de beoordeling van tutoren is over het algemeen van meet af aan al zo hoog dat verdere groei bijna onmogelijk is. Deze verklaring is weinig aannemelijk zoals we dadelijk zullen zien in tabel 5.11. De gemiddelde scores op het merendeel van de items laat van begin af aan voldoende ruimte om eventuele stijgingen te meten.
Een tweede, meer aannemelijke, verklaring is dat blijkbaar geen leereffecten optreden. Ervaren tutoren functioneren niet anders dan onervaren tutoren. Een laatste verklaring is dat de feedback die door de projectgroep programma-evaluatie verschaft wordt, blijkbaar geen effect heeft op tutoren. Ook deze verklaring lijkt aannemelijk omdat immers uit onderzoek gebleken is dat schriftelijke feedback alleen effect heeft als deze gepaard gaat met persoonlijk contact tussen evaluator en docent (Cohen, 1980). We komen op de laatste twee verklaringen terug bij de behandeling van de vraag of verschillen tussen tutoren consistent zijn.
Twee correlaties blijken in tabel 5.10 statistisch significant te zijn namelijk: de negatieve correlatie tussen de ervaringsvariabele en vraag 35 ("De tutor gaf de indruk zijn/haar rol plezierig te vinden") binnen de groep tutoren die 10 tutorschappen vervuld hebben, en de eveneens negatieve correlatie tussen de ervaringsvariabele en vraag 38 ("De tutor stuurde regelmatig met zijn eigen vakkennis de discussie"). De eerste genoemde correlatie suggereert dat binnen subgroep 10 de tutoren na verloop van tijd minder plezierig aan hun rol beleefden. Op zich lijkt dit een aannemelijke uitkomst, omdat men zou kunnen verwachten dat docenten na verloop van tijd uitgekeken raken op de tutorrol. Deze interpretatie wordt echter verzwakt door het feit dat slechts zes tutoren in deze groep zaten en doordat het verschijnsel niet optreedt bij groepen die negen of minder tutorschappen vervuld hebben. De vraag rijst derhalve in hoeverre dit resultaat gevonden zou worden bij grotere aantallen tutoren die een reeks van 10 tutorschappen vervuld hebben. De negatieve correlatie tussen vraag 38 en de ervaringsvariabele suggereert dat bij het toenemen van het aantal tutorschappen, bijsturing van de discussie op basis van eigen vakkennis afneemt. Een interpretatie van deze uitkomst luidt dat tutoren na verloop van tijd een zekere terughoudendheid gaan vertonen om discussies op basis van eigen vakkennis bij te sturen. Een andere interpretatie is dat tutoren steeds meer ervaring krijgen in het begeleiden van onderwijsgroepen, waarbij vooral gebruikt gemaakt wordt van groepsdynamische vaardigheden die het groepsproces stimuleren. In deze visie wordt door tutoren minder een beroep gedaan op vakkennis als middel om de discussie te stimuleren. Beide interpretaties

vinden echter onvoldoende rechtvaardiging in de gegevens van tabel 5.10. In de eerste plaats is de correlatie, hoewel significant, laag. In de tweede plaats zouden beide interpretaties ruggesteun moeten vinden in de correlaties tussen de andere vragen en de ervaringsvariabele.

Uit de hoogte van de correlatie in tabel 5.11 blijkt, dat er geen substantiële redenen zijn om stochasticiteit in de reeks uit te sluiten. Een verklaring voor het ontbreken van significante correlaties tussen items en de ervaringsvariabele zou, zoals gezegd, het optreden van "ceiling effecten" kunnen zijn. In onderstaande tabel 5.11 zijn per tijdstip de gemiddelde scores van alle tutoren weergegeven. Deze tabel is gebaseerd op het hiervoor beschreven "tweede design" waarmee vergelijkingen tussen tijdstippen, onafhankelijk van het aantal tutorschappen van tutoren, gemaakt kunnen worden. Het symbool N staat in deze tabel voor het aantal tutoren dat op het betreffende tijdstip beoordeeld werd. Uit deze tabel blijkt dat hooguit bij de items v33, v34 en v44 sprake zou kunnen zijn van dergelijke effecten. Bij de andere items is dit onwaarschijnlijk. Zoals te verwachten, viel gegeven de niet-significante correlaties tussen de ervaringsvariabele en de scores op de items, blijken de scores binnen een item over een reeks tijdstippen, vrij stabiel zijn.

Tabel 5.11: Gemiddelde scores per tijdstip, berekend over alle tutoren.

	Tutorschap									
Vraag	1	2	3	4	5	6	7	8	9	10
v33	4.0	4.1	4.1	4.2	4.0	4.0	4.2	4.1	4.1	4.1
v34	4.1	4.2	4.2	4.3	4.1	4.2	4.4	4.2	3.8	4.3
v35	4.1	4.1	4.1	4.1	4.0	3.9	4.1	4.0	3.6	4.0
v36	3.5	3.6	3.6	3.6	3.4	3.4	3.7	3.6	3.3	3.7
v37	3.8	3.8	3.9	3.9	3.7	3.7	3.8	3.6	3.7	4.0
v38	3.1	3.0	3.1	3.0	2.9	2.7	2.7	2.7	3.1	3.1
v39	3.5	3.5	3.6	3.6	3.4	3.5	3.7	3.5	3.8	3.8
v40	2.7	2.7	2.7	2.8	2.7	2.8	2.9	2.9	2.7	2.9
v41	3.0	3.0	3.1	3.0	3.0	2.9	3.0	2.7	2.3	3.5
v42	2.7	2.7	2.8	2.8	2.8	2.7	2.8	2.5	2.4	2.9
v43	3.6	3.6	3.6	3.6	3.6	3.8	3.7	3.6	3.6	3.9
v44	4.1	4.1	4.1	4.2	4.0	4.0	4.2	4.1	3.6	4.1
N	326	217	161	111	81	49	29	14	8	6

5.6.8 Conclusies.

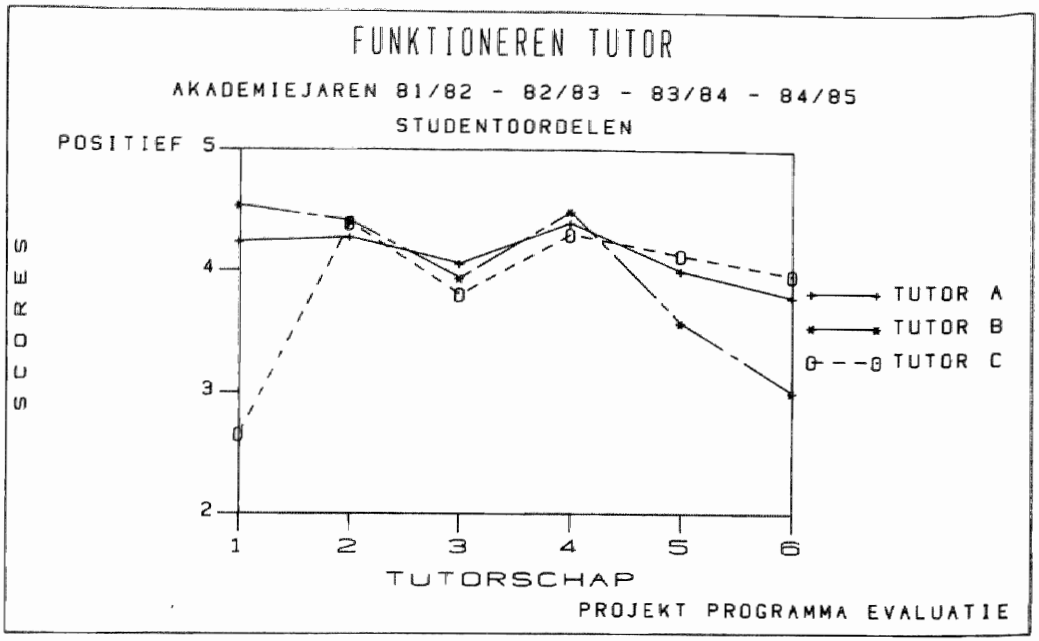
Uit het onderzoek naar de stabiliteit van tutorgedrag binnen een bepaalde tijdsperiode blijkt, dat na verloop van tijd geen stijgende of dalende tendens wordt waargenomen. De tijdreeks vertoont een stochastisch patroon. Deze gegevens leiden tot de voorlopige conclusie dat ervaring blijkbaar geen rol speelt in het functioneren van tutoren, althans voor zover gemeten met studentoordelen op, de vragenlijst voor de

evaluatie van het medisch onderwijs. Een andere voorlopige conclusie is dat feedback met studentoordelen op, geen rol lijkt te spelen bij het verbeteren van het functioneren van tutoren.

5.6.9 Onderzoeksvraag 2:
Zijn er consistente verschillen tussen tutoren te ontdekken? (Zijn tutoren in te delen naar "goede" en "slechte" ?).

In het eerste onderzoek is geconstateerd dat de tijdreeksen betreffende tutorbeoordelingen een stochastisch karakter vertonen. Een voor de hand liggende vraag is dan of verschillen tussen tutoren wel stabiel zullen zijn. Immers, zoals we in de inleiding van deze studie zagen, kan de vraag naar stabiliteit van docentengedrag op twee manieren opgevat worden. In het onderstaande gedeelte worden de resultaten uit het onderzoek naar stabiliteit van verschillen tussen tutoren bediscussieerd.
Eerder hebben we al geconstateerd dat uit ander onderzoek gebleken is dat er weinig consistentie is in de onderlinge verschillen tussen docenten (Shavelson & Dempsey-Atwood, 1976). Docentengedrag bleek per dag of zelfs per uur van een dag sterk te kunnen variëren.
Ging het in de vorige studie om de vraag of tutoren na verloop van tijd beter of slechter beoordeeld werden, in deze studie gaat het om de vraag of tutoren stabiel of wisselend gedrag vertonen.
In onderstaande figuur 5.12 is, als voorbeeld van de stabiliteit van tutorgedrag, de beoordeling van drie tutoren op een item van de vragenlijst weergegeven. Het betreft het item "De tutor functioneerde over het geheel genomen goed in zijn/haar rol". De tutorbeoordeling heeft betrekking op een tijdreeks van zes tutorschappen. De tutoren zijn afkomstig uit drie verschillende clusters van vakgebieden. Tutor A is afkomstig uit de sociaal-wetenschappelijke disciplines, tutor B uit de klinische disciplines, en tutor C uit de biomedische disciplines. Op tijdstip 1 werden deze tutoren geselecteerd op de hoogte van hun beoordeling. Tutor B had op dat tijdstip de hoogste score van alle tutoren, tutor C de laagste score en tutor A een redelijk hoge score. Als de verschillen tussen tutoren stabiel zijn dan kan men verwachten dat deze rangorde over een reeks tutorschappen gehandhaafd blijft. Tutor C zou bijvoorbeeld in dat geval de slechtste beoordeelde tutor van de drie tutoren blijven. Uit onderstaande figuur blijkt dat dit niet het geval is. De scores volgen een onvoorspelbaar verloop, met als uitkomst dat degene die in het begin de laagste score kreeg, op het einde de hoogste score had en andersom.

Figuur 5.12 Studentoordelen over drie tutoren, op het item "tutor functioneerde over het geheel genomen goed".



Deze beoordelingen in figuur 5.12 dienden slechts als voorbeeld om te illustreren dat tenminste voor sommige tutoren de stelling niet opgaat dat wie goed dan wel slecht begint ook goed respectievelijk slecht eindigt. De vraag rijst dan ook hoe stabiel de verschillen tussen alle tutoren zijn.
In het onderzoek werden de studentoordelen over de tutoren op achtereenvolgende tijdstippen met elkaar gecorreleerd. Voor de items die betrekking hadden op het functioneren van de tutor, werd een correlatiematrix opgesteld waarin de correlaties opgenomen waren tussen een item op tijdstip 1 tot en met tijdstip 10. De correlaties werden op twee manieren berekend. Volgens de eerste methode werden correlaties berekend op basis van de gemiddelde onderwijsgroepscore van een item. Bij de tweede methode werden deze gemiddelde scores afgerond en aldus teruggebracht tot de oorspronkelijke scores van de items. Op deze wijze werd voorkomen dat kleine fluctuaties in de gemiddelde scores ertoe zou leiden dat de correlaties tussen metingen ten onrechte laag zouden blijven.
Iedere correlatiematrix bestond uit 10(10-1)/2 correlaties tussen metingen op 10 tijdstippen. Per correlatiematrix werd de mediane correlatie berekend. In onderstaande tabel 5.12 is als voorbeeld een dergelijke correlatiematrix voor item v33 weergegeven. De correlaties zijn berekend op basis van niet-

afgeronde gemiddelde scores en kunnen opgevat worden als test-hertest betrouwbaarheden.

Tabel 5.12: Correlaties tussen de scores op V33 over 10 tijdstippen.

tijdstip	1	2	3	4	5	6	7	8	9	10
1	-	(201)	(150)	(106)	(78)	(47)	(24)	(14)	(8)	(6)
2	.20"	-	(145)	(104)	(74)	(47)	(28)	(14)	(8)	(6)
3	.16'	.32"	-	(102)	(73)	(42)	(25)	(12)	(7)	(6)
4	.08	.18'	.24'	-	(78)	(46)	(28)	(13)	(7)	(5)
5	.25'	.28"	.17	.17	-	(48)	(29)	(14)	(8)	(6)
6	.06	-.01	.35'	-.11	.22	-	(29)	(14)	(8)	(6)
7	.25	.06	.38'	.01	.25	.17	-	(14)	(8)	(6)
8	.36	.51'	.28	.36	.19	-.07	-.08	-	(8)	(6)
9	.84"	.43"	-.14	.14	.78'	.63'	-.22	.71'	-	(6)
10	.03	-.72'	.66	.43	-.26	-.18	.05	.10	.03	-

Noot. Correlaties met - " teken - zijn statistisch significant ($p < .01$), correlaties met - ' teken - zijn statistisch significant ($p < .05$). Tussen haakjes staan de aantallen beoordeling vermeld waarop de correlaties gebaseerd zijn.

Uit de gegevens in deze tabel blijkt dat de correlaties variëren tussen -.72 en .84. De mediane correlatie in deze tabel bedraagt .18.
In onderstaande tabel 5.13 zijn de mediane correlaties van de items v33 tot en met v44 volgens beide methodes weergegeven. Bij de tweede methode zijn bovendien de laagste en hoogste correlatie uit de correlatiematrix toegevoegd om een beeld te krijgen van de range waarbinnen de correlaties liggen. Uit de resultaten in tabel 5.13 blijkt dat beide methoden vergelijkbare i.c. lage mediane correlaties (± .20) opleveren. De correlaties berekend met behulp van methode 2 zijn over het geheel genomen iets lager.
Deze resultaten komen sterk overeen met het eerder genoemde onderzoek van Shavelson en Dempsey-Atwood (1976) die mediane correlaties vonden rondom .17. De range van de correlatieco-ëfficiënten blijkt zeer groot te zijn. Bij sommige items liggen de correlaties zelfs tussen -.63 en 1.0, of tussen -1.0 en .77.

Tabel 5.13: Mediane, laagste en hoogste correlatie per item van de correlatiematrix, bestaande uit correlaties tussen tijdstippen binnen een item.

Item	Methode 1 Mediane correlatie	Methode 2 Mediane correlatie	Laagste correlatie	Hoogste correl.
v33	.18	.11	-.40	.59
v34	.18	.07	-.32	.75
v35	.13	.15	-.61	.73
v36	.20	.16	-.63	1.00
v37	.27	.21	-.77	.82
v38	.38	.34	-.79	.77
v39	.29	.30	-.09	.79
v40	.22	.10	-.60	.59
v41	.17	.17	-.49	.77
v42	.14	.10	-1.00	.77
v43	.36	.29	-.84	.78
v44	.21	.18	-.54	.75

De resultaten in tabel 5.13 lijken tenminste één conclusie te rechtvaardigen, namelijk dat individuele verschillen tussen tutoren niet stabiel zijn. Deze resultaten bevestigen het beeld dat eerder in figuur 5.12 geschetst werd van de beoordeling van drie tutoren.

5.6.10 Conclusies.

In de inleiding van deze studie werden een aantal bestaande maatregelen van de medische faculteit genoemd om de kwaliteit van tutoren te bewaken: verplichte cursussen en het geven van feedback aan tutoren. Bovendien werd gewezen op maatregelen die de faculteit zou kunnen nemen ten aanzien van slecht functionerende tutoren: een verplichte herhaaltraining, of het uitsluiten van slecht functionerende tutoren van verdere tutorschappen.
In studie 4 is onderzocht in hoeverre dergelijke maatregelen gebaseerd zouden kunnen worden op evaluatiegegevens over tutoren. Uit de onderzoeken in deze studie blijkt dat er geen sprake is van stabiel tutorgedrag over een bepaalde periode, en dat er geen sprake is van consistentie in verschillen tussen tutoren. Beide resultaten bevatten aanwijzingen dat het instellen van verplichte herhaaltrainingen voor slecht functionerende tutoren, of het uitsluiten van deze tutoren voor verdere tutorschappen, zinloos is. Immers, de resultaten lijken de conclusie te rechtvaardigen dat er geen permanent slecht functionerende tutoren bestaan. De resultaten lijken eveneens de conclusie te rechtvaardigen dat feedback in de huidige vorm geen of nauwelijks effect heeft op tutoren. Als de feedback naar tutoren effect zou hebben dan zou men na

verloop van tijd betere tutorbeoordelingen mogen verwachten. Dit bleek niet het geval.
Bij beide conclusies dient een kanttekening gemaakt te worden. In het hier beschreven onderzoek vervulden docenten relatief weinig tutorschappen per jaar, gemiddeld niet meer dan twee. De vraag rijst of dezelfde resultaten gevonden zouden zijn als docenten meer dan twee tutorschappen per jaar zouden vervullen, waardoor opeenvolgende perioden waarin ervaring zou kunnen worden opgedaan, dichter bij elkaar zouden liggen. De kans bestaat immers dat tutoren datgene wat eerder geleerd is gewoonweg vergeten.

5.7 Conclusies m.b.t. de bruikbaarheid van studentoordelen.

De belangrijkste doelstellingen van het evaluatiesysteem van de medische faculteit is dat het docenten en bestuurders van bruikbare informatie voorziet die zou kunnen bijdragen tot beheersing c.q. verbetering van de kwaliteit van het onderwijs. In de inleiding van dit hoofdstuk werd opgemerkt dat twee factoren een rol spelen bij de realisatie van deze doelstelling: 1) de betrouwbaarheid en validiteit van de gegevens en 2) de bruikbaarheid van de gegevens.
In dit hoofdstuk is getracht om, aan de hand van vier studies, te illustreren op welke wijze de evaluatiegegevens gebruikt kunnen worden en hoe zij in het verleden gebruikt zijn.
Twee vragen stonden in deze studies centraal: de vraag naar de wijze waarop van die informatie gebruikt gemaakt kan worden, en de vraag naar de effecten van de evaluatie op het onderwijs. Uit de studies bleek dat een effectief gebruik van evaluatiegegevens beïnvloed wordt door een aantal factoren: de mate van afstemming van het onderwijsbeleid op de onderwijsevaluatie, de mate van contact tussen evaluator en docent, en het klimaat waarin de evaluatie plaatsvindt (de betrokkenheid van docenten of bestuurders).
Bovendien bleek uit de studies dat het evaluatiesysteem een goed instrument voor kwaliteitsbewaking biedt. Veranderingen in het onderwijs werden duidelijk zichtbaar in de evaluatieresultaten.

Summary.

In this dissertation an approach to program evaluation is described which is used at the medical faculty of the University of Limburg, the Netherlands. The main features of this approach are its reliance on comparative data with respect to the courses to be evaluated and its use of both quantitative and qualitative information. The main topics of this dissertation are the measurement and improvement of the quality of the educational program of the medical faculty. The reliability, validity and usefulness of this evaluation approach is extensively reviewed and investigated.

In chapter 1 a review is given of recent developments of evaluation in higher education. Over the last decade educational evaluation has assumed increasing importance. The demand for accountability in education has shifted from broad issues of finance and program management to specific concerns about the quality of education and teachers. These concerns have led to a resurgence of interest in evaluating teachers and the development of new systems for teacher evaluation.

Chapter 1 describes several evaluation approaches which measure different aspects of teaching and the teacher. They rely on different conceptions of what demonstrates quality and on diverse notions of how to measure quality. These evaluation approaches are discussed with reference to their theoretical underpinnings and to the organizational context they are to be used. It seems that there is a tendency in educational evaluation to move away from measurement of the effectiveness of teaching in terms of end results and towards emphasis on processes; towards unintended outcomes as well as those which are intended and enshrined in the objectives. Approaches which are based on the concept of goal-oriented evaluation have lost importance because they focus too sharply upon intentions and not enough on what actually happens. On the other hand approaches like illuminative evaluation have gained increasing interest. These approaches pay more attention to the perceptions, opinions and behaviour of staff and students with a variety of research methods.

Perhaps the most widely used method of looking at the process of education is student questionnaires on lecturing. Students are asked to provide a number of ratings on their instructors and courses. Student ratings often play a dual role in educational evaluation. If the evaluations are to be used for formative diagnostic purposes, then elaborate refinement of the evaluation instrument seems not necessary and questions concerning teachers' style, text and examinations are all suitable. On the other hand, for purposes of policy making a standard questionnaire which has undergone considerable refinement is necessary, and only items covering aspects of good teaching should be used. Norms are needed against which numerical ratings can be compared. Furthermore, an equitable process is needed for communicating the results of evaluati-

on.
Ultimately the succes or failure of an evaluation approach depends upon whether or not educational programs change in the way desired by decision-makers, teachers or evaluators. Unfortunately the relationship between evaluation and improvement of educational quality is highly problematic. Notoriously, studies of the utilization of evaluation results show that evaluation does not seem to have an impact on the improvement of educational quality. Several variables which seem to affect utilization are described in chapter 1.

In chapter 2 the evaluation system of the medical faculty is discussed extensively. The evaluation procedure that has been developed consists of a general screening of all blocks carried out with the help of questionnaires. Instructors make up specific questions for their courses while 60 standard items are included for both formative and summative purposes. As quickly as possible a short evaluation report is written on each block, presenting the findings on both standard and specific questions and presenting the interpretations of the evaluator based on an analysis of the block-book, informal discussions with students and tutors. The report also contains suggestions for possible improvement of the block-book. Chapter 2 gives a description of the standard items of the questionnaire. These items are derived from theories on problem-based learning, theories of school learning and on opinions of teachers and student about features of effective problem-based learning.
Chapter 3 provides an extensive review on the reliability and validity of student ratings of instruction. In many studies of college students' ratings of their teachers and their courses, the degree of internal consistency among students is taken as a mean for establishing the interrater reliability of student raitngs. Research findings in general suggest that reliabilities of individual ratings or the single rater, consistency among students in their ratings is moderate, at best. By contrast, under the assumption that students are independent replicates (except for random error), these modest interrater associations do produce substantial reliabilities when the ratings of students (10 or more) are averaged together, thus indicating that the average or composite ratings of teachers and courses are dependable measures.
Perhaps the most critical question about student ratings of instruction is whether they are valid: whether they actually measure teaching effectiveness. Student ratings are only one measure of teaching effectiveness and are difficult to validate since no universal criterion of effective teaching exists. Consequently several approaches are used to validate student ratings. For example, a typical approach consists of validating students' evaluations against student learning, measured by objective examination. Although seperate research findings have not always been consistent, findings from meta-

analyses suggest that student ratings correlate moderately with measures of student learning. Therefore, student ratings of instruction may be considered as a valid index for instructional effectiveness. Another and more sophisticated approach for validating student ratings consists of multi-trait-multimethod analysis. Within this approach it is tried to demonstrate that student ratings are logically related to various other indicators of effective teaching. Rating factors are required to be most highly correlated with variables to which they are most logically and theoretically related and less or not correlated with other variables. Multitrait-multimethod studies showed that the construct validity of student ratings was supported by research findings. Students' evaluations showed for instance significant agreement with instructor self-evaluations, to a degree that is acceptable to most researchers.

In chapter 4 six empirical studies are described on the reliability and validity of student ratings. These ratings were obtained with the evaluation instrument as described in chapter 2.

In the first study a factor analysis was done to test the multidimensionality of student ratings. Fifteen components were identified in the analysis which gave evidence for the multidimensionality of student ratings. Furthermore the results demonstrated that student ratings were nearly not affected by halo-effects.

The second study investigated the relative agreement among different items designed to measure the same factor. The coefficient alphas for the different evaluation factors is about .80. Therefore it may be concluded that the internal consistency among items is high.

The third study investigated the reliability of students' evaluations. According to one conceptualisation of reliability called the intraclass correlation, a reliable item is one in which there is agreement among ratings within each tutorial group, but consistent differences between the ratings of diffent classes. Research findings showed that the average reliability is about .70 when based upon a response of 7 students, but falls to .25 when based upon one. Given a sufficient number of students, the reliability of class-average student ratings compares favorably with that of objective tests.

Student ratings were validated in the fourth study against student learning as measured by block tests. Examination scores were calculated as percentages-correct scores in order to make comparisons possible between courses. Correlations between testscores and student ratings were calculated for four seperate academic years. The research findings suggested that substantial correlations exist within the first academic year which was under scope of study. The other academic years showed controversial results. Nearly no significant correlations between students evaluations and student learning were found. It was concluded that these results may be confounded

by changes in the examination system of the medical faculty. Lack of absolute control over the conditions of measurement suggest the need for caution in the interpretation of these results; on the other hand, this very absence of control may make this study a more valid test of real-life evaluation procedures.

A construct validation approach was followed in the fifth study to assess the convergent and divergent validity of student ratings. Convergence is in multitrait-multimethod studies inferred from the magnitude of agreement between diffent methods of assessing the same evaluation dimensions. Convergent validity was demontrated for 5 of 7 factors. Convincing evidence was also found for divergent validity. In the final study of chapter 4 the generalizability of the multidimensional structure of students' ratings was examined. The invariance of correlation matrices among factors of the questionnaire, across different academic years, was studied. The research findings demonstrated the invariance of the multivariate structure of the ratings.

Taken together, the results of these studies provide strong evidence for the reliability and validity of student ratings. Finally, in chapter 5 the usefulness of student ratings was examined. The extent to which evaluation findings are utilized is issue of continual concern in evaluation research. With regard to the reliability and validity students' ratings, the situation is quite satisfactory. It seems unclear, however, whether the quality of teaching and education actually improves as a result of evaluation research. Therefore it is necessary to find out if, and under what conditions, students' ratings are useful as feedback to teachers and decisionmakers.

In chapter 5, four studies are described which deal with questions about the usefulness of student ratings. The first two studies may be characterized as case-studies. These studies provide cases of the effects of students' ratings on the quality of problem-based learning. These studies were conducted in settings of the third year of the educational program of the medical faculty. The main conclusion from these studies is that students' ratings do seem effective for the purpose of improving educational quality. These two studies suggest that feedback, coupled with a discussion with a consultant, can be an effective intervention for the improvement of educational quality.

In the third study an experiment around the organisation of planninggroups is described. In this experiment tutors got more responsibility, than usual, in the process of decision-making around courses. The question was whether this change in the organisational structure of the faculty might influence the quality of teaching and education. Student ratings are used to provide a basis for informed administrative decisions. This study provides empirical evidence of improving teaching quality resulting from the experimental changes. Therefore it is recognized that student ratings are

a valid and useful indicator to measure change. Finally, this study shows that student ratings might be considered as a useful indicator for evaluating the merits of a program.
The fourth and final study was conducted to answer the question whether tutor behaviour is stable. The purpose for assessing the stability of tutor behaviour was twofold. The first question was whether student ratings of tutor behaviour are representative of the tutor's usual and customary way of behaving. Because almost no empirical research exists to judge the presumption of consistency, this study might provide insight in the question whether tutors consistently behave in a particular way. An affirmative answer might, for example, indicate directions for faculty's policy towards tutors by whom undesirable behaviour is observed. The second question is concerned with the stability of behaviour over time. This issue is particular important to answer questions about the effects of student ratings the possible improvement of teaching. The research findings demonstrated that tutor behaviour is not consistent. Depending on the situation, tutors change their behaviour in particular directions. It was also demonstrated that tutors don't improve the quality of teaching over time. This result suggests that feedback (consisting of written reports only) towards tutors, does not produce substantial and significant positive changes.

Bijlage 1. Tekstuele veranderingen in de vragenlijst voor studenten.

Academiejaar			
84/85	83/84	82/83	81/82
Vraagnummer			
1 Over het geheel genomen heb ik de afgelopen periode prettig gewerkt			
2 Het blok sloot goed aan op mijn voorkennis			
3 De doelstellingen van dit blok waren mij duidelijk			
4 Het programma vergde veel studietijd			niet opgenomen
5 De onderwerpen in dit blok waren volgens mij nuttig in het kader van de medische studie			De onderwerpen die in dit blok aan de orde zijn gekomen heb ik in het kader van mijn studie relevant ervaren
6 Het programma van dit blok heeft mijn attitude t.a.v. de gezondheidszorg beïnvloed			
7 De leerstof van dit blok was moeilijk			Vergeleken met de vorige blokken die ik gevolgd heb was de leerstof van dit blok moeilijk
8 Ik heb in dit blok veel opgestoken			niet opgenomen
9 Ik vond de in dit blok aangeboden leerstof interessant	niet opgenomen	niet opgenomen	niet opgenomen
10 Informatie m.b.t. opzet en werkwijze was duidelijk gepresenteerd			
11 De taken waren voor het merendeel duidelijk omschreven			De taken waren voor het merendeel voldoende duidelijk geformuleerd
12 De taken leenden zich voor het merendeel voor een systematische aanpak			De taken leenden zich voor het merendeel voor een gestructureerde aanpak
13 De taken waren zo sterk gestructureerd dat er weinig lol aan te beleven was			De taken waren zo voorgekookt dat er weinig lol aan te beleven was
14 De taken gaven voldoende aanleiding tot een zinnige groepsdiscussie			De taken stimuleerden de groepsdiscussie
15 De taken gaven voldoende aanleiding tot zelfstudie			
16 Ik heb in dit blok in belangrijke mate onafhankelijk van het blokboek gestudeerd			niet opgenomen

Academiejaar			
84/85	83/84	82/83	81/82
Vraagnummer			
17 De taken gaven voldoende aanknopingspunten voor het formuleren van leerdoelen			niet opgenomen
18 Er was een grote variëteit aan onderwerpen in het blokboek			niet opgenomen
19 Er was een grote variëteit aan taken in het blokboek			
20 Er was een grote variëteit aan hulpmiddelen bij de taken in het blokboek			niet opgenomen
21 De taken waren aanleiding om onderwerpen uit basisvakken te bestuderen			niet opgenomen
22 Door het werken met de taken in het blokboek was het mogelijk de blokdoelstellingen te realiseren			niet opgenomen
23 De hoeveelheid taken in het blokboek moet worden uitgebreid			
24 De onderwijsgroep maakte gebruik van systematische werkprocedures bij het aanpakken van de taken			De groep maakte gebruik van duidelijke werkprocedures bij het aanpakken van problemen
25 Het werken in de groep betekende een stimulans voor mijn zelfstudie-activiteiten			
26 In de onderwijsgroep werden steeds duidelijke afspraken gemaakt m.b.t. de te bestuderen stof			In de onderwijsgroep werden steeds duidelijke afspraken gemaakt m.b.t. de te ondernemen studieaktiviteiten
27 Iedereen hield zich aan zijn afspraken			
28 Ik heb de bijeenkomsten als prettig ervaren			niet opgenomen
29 De bijeenkomsten waren productief			
30 Iedereen leverde een aktieve bijdrage			
31 Ik heb de onderwijsgroepsbijeenkomsten als een rem ervaren in de voortgang van mijn studie			niet opgenomen

Academiejaar Vraagnummer	84/85	83/84	82/83	81/82
32 Ik heb in dit blok in belanrijke mate onafhankelijk van de leerdoelen van de onderwijsgroep gestudeerd		Niet opgenomen	niet opgenomen	niet opgenomen
33 De tutor gaf blijk een goed begrip te bezitten van de doelstellingen van het blok				
34 De tutor leek op de hoogte van de onderwijskundige uitgangspunten van het onderwijssysteem				
35 De tutor gaf de indruk zijn/haar rol plezierig te vinden				
36 De tutor stimuleerde tot hard werken				
37 De tutor stelde regelmatig discussiestimulerende vragen				
38 De tutor stuurde regelmatig met zijn eigen vakkennis de discussie				
39 De tutor stimuleerde het maken van afspraken mbt de te bestuderen stof				De tutor lette erop dat afspraken mbt de te bestuderen stof gemaakt werden
40 De tutor controleerde het nakomen van gemaakte afspraken				De tutor controleerde of afpsraken ook nagekomen werden
41 De tutor stimuleerde het raadplegen van inhoudsdeskundigen				De tutor stimuleerde het raadplegen van inhoudsdeskundigen en het gebruikmaken van andere leer- en evaluatiemiddelen
42 De tutor stimuleerde het gebruik van andere leer- en evaluatiemiddelen				
43 Regelmatige evalueerde de tutor met ons de gang van zaken in de onderwijsgroep				
44 De tutor funktioneerde over het geheel genomen goed in zijn/haar rol als tutor				De tutor functioneerde alles bij elkaar genomen goed in zijn/haar rol van tutor
45 Ik vond de training(en) lichamelijk onderzoek in dit blok zinvol (incl. patiëntencontacten)			niet opgenomen	niet opgenomen

Academiejaar

Vraagnummer	84/85	83/84	82/83	81/82
46	Ik vond de training(en) therapeutische vaardigheden in dit blok zinvol		niet opgenomen	niet opgenomen
47	Ik vond de training(en) laboratoriumvaardigheden in dit blok zinvol		niet opgenomen	niet opgenomen
48	Ik ben tevreden over de begeleiding bij bovengenoemde trainingen		niet opgenomen	niet opgenomen
49	Ik vond de training(en) sociale vaardigheden in dit blok zinvol		niet opgenomen	niet opgenomen
50	Ik vond de simulatie-patiënten-kontakten in dit blok zinvol		niet opgenomen	niet opgenomen
51	Ik ben tevreden over de begeleiding bij de training (en) sociale vaardigheden c.q. simulatie patiënten-nabespreking		niet opgenomen	niet opgenomen
52	In het studielandschap was een voldoende verscheidenheid aan literatuur beschikbaar			
53	Er waren voldoende AV-middelen beschikbaar			
54	De beschikbare AV-middelen waren inhoudelijk van goede kwaliteit			niet opgenomen
55	Onze onderwijsgroep heeft een aantal malen een inhoudsdeskundige geraadpleegd, n.l.			Mijn groep heeft meer dan één keer pogingen gedaan om een inhoudsdeskundige te bereiken
56	De bloktoets sloot aan bij de door mij bestudeerde onderwerpen			De formatieve evaluatie sloot aan bij door mij bestudeerde onderwerpen
57	De bloktoets toetste onderwerpen die in de doelstellingen			niet opgenomen
58	De zelfevaluatiemiddelen sloten aan bij de inhoud van het blok			niet opgenomen
59	Hoeveel tijd hebt u, gemiddeld genomen, per week aan literatuurstudie besteed			Hoeveel tijd hebt u, gemiddeld genomen, per week aan individuele studie besteed
60	Als u het totale onderwijs zoals u het in dit blok meemaakte, globaal genomen, een cijfer zou moeten geven op een schaal van 1 tot 10 (6 is een voldoende), welk cijfer geeft u dan?			

Bijlage 2: TUTORVRAGENLIJST (academiejaar 1984/1985).

BLOKNUMMER: .....
ONDERWIJSGROEPSNUMMER: .....
NAAM TUTOR: .....

VAK TUTOR:
1 *KLINISCH*
2 *BIOMEDISCH*
3 *GEDRAGSWETENSCHAPPELIJK*
4 *ANDERE*
(AANKRUISEN WAT VAN TOEPASSING IS)

		Volledig oneens				*Volledig eens*
	ALGEMENE INDRUK.					
1.	*Over het geheel genomen hebben de studenten de afgelopen periode plezierig gewerkt.*	1	2	3	4	5
2.	*Ik heb de indruk dat dit blok goed aansloot op hun voorkennis.*	1	2	3	4	5
3.	*De doelstellingen van dit blok waren hun duidelijk.*	1	2	3	4	5
4.	*Volgens de studenten vergde het programma veel studietijd.*	1	2	3	4	5
5.	*De onderwerpen die in dit blok aan de orde zijn gekomen hebben de studenten als relevant ervaren.*	1	2	3	4	5
	HET BLOKBOEK.					
6.	*Informatie m.b.t. opzet en werkwijze was voldoende duidelijk gepresenteerd.*	1	2	3	4	5
7.	*De taken waren voor het merendeel duidelijk omschreven.*	1	2	3	4	5
8.	*De taken leenden zich voor het merendeel voor een systematische aanpak.*	1	2	3	4	5
9.	*De taken waren zo sterk gestructureerd dat er weinig lol aan te beleven was.*	1	2	3	4	5
10.	*De taken gaven voldoende aanleiding tot een zinnige groepsdiscussie.*	1	2	3	4	5
11.	*De taken gaven voldoende aanleiding tot zelfstudie.*	1	2	3	4	5
12.	*Ik heb de indruk dat de studenten in dit blok in belangrijke mate onafhankelijk van het blokboek gestudeerd hebben.*	1	2	3	4	5
13.	*De taken gaven voldoende aanknopingspunten voor het formuleren van leerdoelen.*	1	2	3	4	5
14.	*Er was een grote variëteit aan onderwerpen in het blokboek.*	1	2	3	4	5
15.	*Er was een grote variëteit aan taken in het blokboek.*	1	2	3	4	5

16.	Er was een grote variëteit aan hulpmiddelen bij de taken in het blokboek.	1	2	3	4	5
17.	De taken waren aanleiding om onderwerpen uit basisvakken te bestuderen (anatomie, biochemie, etc.)	1	2	3	4	5
18.	Door het werken met de taken in het blokboek was het mogelijk de blokdoelstellingen te realiseren.	1	2	3	4	5
19.	De hoeveelheid taken in het blokboek moet worden uitgebreid.	1	2	3	4	5

ONDERWIJSGROEP.

21.	De onderwijsgroep maakte gebruik van systematische werkprocedures bij het aanpakken van de taken.	1	2	3	4	5
22.	In de onderwijsgroep werden steeds duidelijke afspraken gemaakt m.b.t. de te bestuderen stof.	1	2	3	4	5
23.	Iedereen hield zich aan zijn afspraken.	1	2	3	4	5
24.	De bijeenkomsten waren productief.	1	2	3	4	5
25.	Iedereen leverde een aktieve bijdrage.	1	2	3	4	5
26.	Ik vond het prettig om met deze groep te werken.	1	2	3	4	5
27.	De studenten waren niet erg mededeelzaam wat betreft hun ervaringen met dit blok t.o.v. mij.	1	2	3	4	5
28.	De studenten hebben in dit blok in belangrijke mate onafhankelijk van de leerdoelen van de onderwijsgroep gestudeerd.	1	2	3	4	5

TUTOR.

29.	Ik was op de hoogte van de onderwijskundige uitgangspunten van het onderwijssysteem.	1	2	3	4	5
30.	Ik had een goed begrip van de concrete doelstellingen van het betreffende blok.	1	2	3	4	5
31.	Ik stelde regelmatig discussie-stimulerende vragen.	1	2	3	4	5
32.	Dit was volgens mij in deze groep noodzakelijk.	1	2	3	4	5
33.	De studenten hadden waardering voor deze inbreng.	1	2	3	4	5
34.	Ik stuurde regelmatig de discussie met mijn eisen vakdeskundigheid bij.	1	2	3	4	5
35.	Dit was volgens mij in deze groep noodzakelijk.	1	2	3	4	5
36.	De studenten hadden waardering voor deze inbreng.	1	2	3	4	5

37. Ik heb er regelmatig op moeten letten dat afspraken m.b.t. de te bestuderen stof gemaakt en nagekomen werden.	1	2	3	4	5
38. Ik nam regelmatig het initiatief om de gang van zaken in de onderwijsgroep te evalueren.	1	2	3	4	5
39. Ik denk dat ik in deze onderwijsgroep nuttig werk heb gedaan.	1	2	3	4	5
41. Ik heb vaak in moeten grijpen als het gesprek te rommelig werd.	1	2	3	4	5
42. De planningsgroep heeft mij vooraf voldoende geïnformeerd m.b.t. mijn taak.	1	2	3	4	5
43. Ik vind dat ik over het geheel genomen als tutor goed gefunctioneerd heb.	1	2	3	4	5

44. Dit blok heeft mij als tutor in totaal ..... uren voorbereidingstijd gekost (onderwijsgroepsbijeenkomsten niet meegeteld)

INHOUDSDESKUNDIGEN

45. Onze onderwijsgroep heeft een aantal malen een inhoudsdeskundige geraadpleegd, namelijk: AANKRUISEN HETGEEN VAN TOEPASSING IS	0	1-2	3-4	5-6	meer dan 6 keer

ENKELE AFSLUITENDE OPEN VRAGEN

46. Welke aspecten of onderdelen van dit blok zouden veranderd moeten worden?

47. Welke aspecten of onderdelen van dit blok vond U erg goed?

48. Heeft U nog op- of aanmerkingen die U in deze vragenlijst nog niet hebt kunnen spuien?

(c)
Vakgroep Onderwijsontwikkeling en
Onderwijsresearch.

Bijlage 3: studie 6, hoofdstuk 4.

Correlatiematrix 1984/1985. N = 357

Schalen	F1	F2	F3	F5	F6	F7	F8	F10	F12	F13	F14	F15
F1	--											
F2	28	--										
F3	48'	42'	--									
F5	55'	14'	36'	--								
F6	37'	11	22'	38'	--							
F7	37'	11	23'	45'	36'	--						
F8	54'	17'	61'	50'	27'	33'	--					
F10	10	35'	28'	02	00	04	18'	--				
F12	-21'	-17'	-37'	-13'	-10	-07	-21	-02	--			
F13	01	20'	11	14'	11	07	10	11	00	--		
F14	48'	19'	25'	47'	37'	38'	30'	07	-13'	03	--	
F15	10	63'	20'	09	06	11	10	40'	-09	23'	10	--

Correlatiematrix 1983/1984. N = 295

Schalen	F1	F2	F3	F5	F6	F7	F8	F10	F12	F13	F14	F15
F1	--											
F2	27'	--										
F3	50'	45'	--									
F5	55'	19'	37'	--								
F6	30'	13	17'	27'	--							
F7	18'	14'	12	20'	15'	--						
F8	51'	26'	56'	37'	18'	09	--					
F10	15'	45'	29'	08	09	09	25'	--				
F12	-25'	-22'	-25'	-03	-13	-06	-25'	-08	--			
F13	-03	14'	-01	-04	05	-12	00	11	01	--		
F14	34'	19'	37'	41'	29'	22'	41'	10	-14'	-04	--	
F15	15'	66'	19'	06	08	09	12	47'	-12	25'	10	--

Correlatiematrix 1982/1983. N = 238

Schalen	F1	F2	F3	F5	F6	F7	F8	F10	F12	F13	F14	F15
F1	--											
F2	26'	--										
F3	56'	42'	--									
F5	66'	16'	46'	--								
F6	43'	18'	32'	39'	--							
F7	06	-01	04	-01	04	--						
F8	70'	23'	68'	59'	39'	06	--					
F10	18'	46'	30'	09	05	-06	17'	--				
F12	-23'	-05	-22'	-16'	-13	-10	-24'	-04	--			
F13	-23'	-12	-17'	-09	-03	-06	-15'	-02	-05	--		
F14	36'	12	26'	38'	05	15'	40'	04	-04	-19'	--	
F15	16'	53'	21'	05	-03	-07	07	42'	04	04	00	--

Noot. Correlaties met - ' teken - zijn statistisch significant ($p < .05$).

Bijlage 4. Hoofdstuk 5, studie 3.

VRAGENLIJST VOOR DE BEOORDELING VAN HET ONDERWIJS IN DE EXPERIMENTELE BLOKKEN VAN DE FACULTEIT DER GENEESKUNDE, DOOR TUTOREN.

Geachte tutor,

Omdat vorig jaar een experiment van start is gegaan om van enkele blokplanningsgroepen de leden tevens in datzelfde blok als tutor te laten fungeren, is er door het project Programma-Evaluatie een extra vragenlijst voor tutoren samengesteld.

De gegevens die door u op deze vragenlijst worden ingevuld, en door het project Programma-Evaluatie worden verzameld en verwerkt, zijn zeer belangrijk voor het eventueel bijstellen van de blokorganisatie.
De resultaten zullen worden gerapporteerd aan alle bij het onderwijs betrokkenen: planningsgroep, Onderwijscommissie FdG, etc.

Enkele opmerkingen over de lijst zelf:
De vragenlijst kent een grote verscheidenheid aan vragen, o.a. meerkeuzevragen, ja/nee vragen en open vragen.
De vragen 17 tot en met 21 bestaan uit beweringen waarop u kunt reageren door omcirkeling van een cijfer.

1 betekent dat u het "volledig oneens" bent met de bewering;
2 betekent "tamelijk oneens";
3 betekent "neutraal", "er tussen in";
4 betekent "tamelijk eens";
5 betekent "volledig eens".

U wordt verzocht de ingevulde vragenlijst zo spoedig mogelijk te sturen aan:
Nelly Plompen
Bureau Onderwijs FdG
secretariaat Programma-Evaluatie
Randwijck.

Wij hopen op uw bereidwillige medewerking.

Wim Gijselaers
Vakgroep Onderwijsontwikkeling en
Onderwijsresearch.

Vragenlijst voor de evaluatie van experimentele blokken.

BLOKNUMMER: .....
ONDERWIJSGROEPSNUMMER: .....
NAAM TUTOR: .....

Algemeen.

1.	*In welk vakgebied bent u thans werkzaam?*	1. *Klinisch* 2. *Bio-medisch* 3. *Sociaal-wetenschappelijk* 4. *Anders, namelijk*
2.	*Sinds wanneer bent u aan de medische faculteit werkzaam?*	*Jaar: .....*
3.	*Heeft u reeds eerder tutorschappen (exclusief dit blok) gehad?*	*Ja / Nee*
4.	*Zo ja, hoeveel*	*1* *2-3* *4-5* *meer*
5.	*Bent u reeds eerder planningsgroeplid geweest?* *Zo ja, in welke blokken, en in welke jaren?* *(jaar = academiejaar)*	*Ja / Nee* *blok ... jaar ..../....* *blok ... jaar ..../....* *blok ... jaar ..../....* *blok ... jaar ..../....*
6.	*Bent u reeds eerder blokcoördinator geweest?* *Zo ja, in welke blokken, en in welke jaren?* *(jaar = academiejaar)*	*Ja / Nee* *blok ... jaar ..../....* *blok ... jaar ..../....* *blok ... jaar ..../....*

Voorbereiding van het blok.

7.	*Hoe vaak is de complete planningsgroep bij elkaar geweest?*	.....

Subgroepen.

8.	*Zijn er in de planningsgroep subgroepen gevormd ter voorbereiding van verschillende onderdelen?* *Zo ja, waarom?*	*Ja / Nee*

9. *Heeft u een zinvolle bijdrage kunnen leveren in deze bijeenkomsten? Zo nee, waarom niet?* — *Ja / Nee*

10. *Welke taken hadden deze subgroepen?*

11. *Hoe vaak heeft u dergelijke bijeenkomsten van subgroepen bijgewoond?* — *aantal: .....*

<u>Planningsgroep.</u>

12. *Heeft u de mogelijkheid gehad om bijdragen van andere leden of subgroepen te beoordelen of te becommentariëren?* — *Ja / Nee*
13. *Stond men open voor elkaars opmerkingen?* — *Ja / Nee*
14. *Welke bijdragen heeft u geleverd in de planningsgroep?*

<u>Bijdrage geleverd aan:</u>	*<u>Welke vorm had deze bijdrage?</u>*
1) Blokopening	*1. Inhoudelijk advies*
	2. Redactioneel/vormtechnisch advies
	3. Uitvoering
2) Taakconstructie	*1. Inhoudelijk advies*
	2. Redactioneel/vormtechnisch advies
	3. Uitvoering
3) Practica	*1. Inhoudelijk advies*
	2. Redactioneel/vormtechnisch advies
	3. Uitvoering
4) Lezingen	*1. Inhoudelijk advies*
	2. Redactioneel/vormtechnisch advies
	3. Uitvoering

5) Literatuurboek	1. Inhoudelijk advies 2. Redactioneel/vormtechnisch advies 3. Uitvoering
6) Tutorinstructie	1. Inhoudelijk advies 2. Redactioneel/vormtechnisch advies 3. Uitvoering
7) Toetsconstructie	1. Inhoudelijk advies 2. Redactioneel/vormtechnisch advies 3. Uitvoering

Tijdsbesteding.

15. Waaraan heeft u de meeste tijd besteed m.b.t. de organisatie/ vormgeving van het blok? (exclusief het tutorschap)

16. Hoeveel tijd heeft u in totaal ongeveer aan deze activiteit besteed?

Oordeel m.b.t. deze organisatievorm.
(Hieronder vindt u enkele uitspraken m.b.t. de docentrol in dit blo

	Volledig oneens			Volledig eens	
17. Ik vond deze manier van werken efficiënt.	1	2	3	4	5
18. In vergelijking met andere blokken voelde ik me in dit blok als tutor meer betrokken.	1	2	3	4	5
19. Ik kon de onderwijsgroep in dit blok, in vergelijking met vorige blokken, beter leiden in de richting van de doelstellingen van het blok.	1	2	3	4	5
20. Ik was, in vergelijking met vorige blokken, beter in staat om gerichte vragen te stellen in de onderwijsgroep.	1	2	3	4	5
21. Ik had als tutor, in vergelijking met eerdere blokken, meer plezier in mijn rol.	1	2	3	4	5

22. *Zou u de onderwijscommissie FdG adviseren meer blokken op deze wijze te organiseren? Zo ja, waarom? Ja / Nee*
Zo nee, waarom?

23. *Heeft u nog op- of aanmerkingen t.a.v. deze wijze van blokorganisatie, die u nog niet eerder kwijt kon in deze vragenlijst?*

Bijlage 5. Voorbeeld tutorbeoordeling. Hoofdstuk 5, studie 4.

PROGRAMMA-EVALUATIE MF BLOK 1.1

ONDERWIJSGROEP *

Number of valid observations (listwise) = 6.00

Variable	Mean	Std Dev	Valid N	Label
VR25	4.6	.5	8	Gebruikmaking system. werkprocedures
VR26	4.1	1.0	8	Stimulans zelfstudie
VR27	3.9	1.1	8	Afspraken werden gemaakt
VR28	3.3	.7	8	Men hield zich aan de afspraken
VR29	4.5	.5	8	Bijeenkomsten prettig
VR30	4.1	.4	8	Bijeenkomsten productief
VR31	3.4	.7	8	Aktieve bijdragen van iedereen
VR32	1.0	.0	7	Bijeenkomsten als rem ervaren
VR33	2.0	.8	7	Onafhankelijk van leer-doelen gestudeerd
VR34	5.0	.0	8	Tutor goed begrip doel-stellingen
VR35	5.0	.0	8	Tutor op hoogte van uitgangspunten
VR36	4.9	.4	8	Tutor vond zijn rol ple-zierig
VR37	4.1	.6	8	Tutor stimuleerde tot hard werken
VR38	4.1	1.1	8	Tutor stelde discus.stim. vragen
VR39	2.6	1.3	8	Tutor stuurde met eigen vakkennis
VR40	3.8	1.2	8	Tutor stimul. het maken van afspraken
VR41	2.8	1.0	8	Controleerde nakomen van afspraken
VR42	4.8	.5	8	Tutor stimul. raadplegen inh.desk.
VR43	4.1	1.0	8	Stimul. gebruikm. van andere leermiddelen
VR44	5.0	.0	8	Tutor evalueerde regel-matig
VR45	5.0	.0	8	Tutor funktioneerde goed

Bijlage 6: Hoofdstuk 5, studie 4.

Per tijdstip werd voor de afzonderlijke items v33 - v44 een eenwegvariantie-analyse (ANOVA) verricht, waarbij getoetst werd of tussen de subgroepen (1-10) statistisch significante verschillen bestonden. De nulhypothese was dat per tijdstip geen verschillen bestonden tussen de subgroepen op de items v33 -v44.
Uit de resultaten bleek dat slechts in drie gevallen de nulhypothese verworpen moest worden. Deze bleken alle op tijdstip 8 voor te komen. Op dit tijdstip bestonden op de items v36, v38 en v42 statistisch significante verschillen tussen de subgroepen 8, 9 en 10. In onderstaande tabellen zijn de resultaten van deze analyse weergegeven.

Tijdstip 8.

Item v36. Variantie-analyse.

	D.F.	SS	MS	F-ratio	F-prob
Tussen Groepen	2	2.40	1.2	6.7	.012
Binnen Groepen.	11	1.95	.18		

Resultaten per groep.

Groep	N	Gem.	Sd
8	6	3.4	.27
9	2	4.6	.00
10	6	3.4	.56

Item v38. Variantie-analyse.

	D.F.	SS	MS	F-ratio	F-prob
Tussen Groepen	2	4.2	2.1	4.9	.03
Binnen Groepen	11	4.7	.43		

Resultaten per groep.

Groep	N	Gem.	Sd
8	6	2.1	.61
9	2	3.5	.63
10	6	3.1	.70

Item v42. Variantie-analyse.

	D.F.	SS	MS	F-ratio	F-prob
Tussen Groepen	2	1.6	.80	4.2	.05
Binnen Groepen.	11	2.1	.20		

Resultaten per groep.

Groep	N	Gem.	Sd
8	6	2.2	.45
9	2	3.2	.07
10	6	2.6	.46

De verschillen tussen de subgroepen op tijdstip 8 lijken voornamelijk te wijten aan subgroep 9. Deze groep bestaat uit 2 tutoren. Gezien het geringe aantal tutoren in deze groep likt de conclusie gerechtvaardigd dat toevalsfactoren waarschijnlijk de verschillen tussen subgroepen verklaren.
Uit het feit dat uit een totaal van 108 variantie-analyses (12 items x 9 tijdstippen) slechts 3 analyses statistisch significante verschillen tussen subgroepen laten zien, kan geconcludeerd worden dat er geen verschillen tussen de groepen lijken te bestaan.

Literatuurverwijzingen.

Abrami, P.C., Leventhal, L., & Perry, R.P. (1979). Can feedback from student ratings help to improve college teaching? *Proceedings of the 5th International Conference on Improving University Teaching*. London.

Abrami, P.C., Leventhal, L., & Perry, R.P. (1982). Educational seduction. *Review of Educational Research, 52*, 446-464.

Academisch Statuut. Algemeen Deel en Memorie van Toelichting. (1981). 's-Gravenhage: Staatsuitgeverij.

Ausubel, D.P. (1968). *Educational psychology: a cognitive view*. New York: Holt, Rinehart & Winston.

Barrows, H.G., & Tamblyn, R.M. (1980). *Problem-based learning* New York: Springer Press.

Bausell, R.B., & Magoon, J. (1972a). Expected grade in a course, grade point average, and student ratings of the course and the instructor. *Educational and Psychological Measurement, 32*, 1013-1023.

Bausell, R.B., & Magoon, J. (1972b). Instructional methods and college student ratings of courses and instructors. *Journal of Experimental Education, 40*, 29-33.

Bausell, R.B., & Magoon, J. (1972c). The persistence of first impressions in course and instructor evaluations. Presented at the Annual meeting of the American Educational Research Association.

Bausell, R.B., Magoon, J. (1972d). *The validation of student ratings of instruction: an institutional research model.* Newark, DE: College of Education, University of Minnesota.

Bendig, A.W. (1953). The relation of level of course achievement to students' instructor course ratings in introductory psychology. *Educational and Psychological Measurement, 13*, 437-448.

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Blom, S.J.M., & Langerak, W.F. (1979). Beoordelen van docenten door studenten. Een literatuurstudie. *Pedagogische Studien, 56,* 308-318.

Bloom, B.S. (Ed.), (1956). *Taxonomy of educational objectives. Handbook I: Cognitive Domain.* New York: Longmans Green.

Bloom, B.S. (1970). Toward a theory of testing which includes measurement -evaluation- assessment. In: Wittrock, M.C., & Wiley, D.E. (Eds.). *The evaluation of instruction.* New York: Holt, Rinehart & Winston.

Bloom, B.S. (1974). Time and learning. *American Psychologist, 29,* 682-688.

Bloom, B.S. (1976). *Human characteristics and school learning.* New York: McGraw-Hill.

Bouhuijs, P.A.J., Gijselaers, W.H., & Kerkhofs, L.M.M. (1984). The use of group data to trace the influence of individual students on group functioning. In: H.G. Schmidt, & Volder, M.L., de, (Ed.) *Tutorials in Problem-Based Learning.*

Brandenburg, D.C., Nassauer, R., & Buckmaster, A. (1981). *Using results from student ratings of instruction: a survey of faculty perception.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Braskamp, L.A., Brandenburg, D.C., & Ory, J.C. (1984). *Evaluating Teaching Effectiveness. A Practical Guide.* Beverly Hills: Sage Publications.

Braskamp, L.A., Caulley, D., & Costin, F. (1979). Student ratings and instructor self-ratings and their relationship to student acheivement. *American Educational Research Journal, 16,* 295-306.

Bruner, J.S. (1961) The act of discovery. *Harvard Educational Review, 31,* 21-32.

Bruner, J.S. (1966). *Towards a theory of instruction.* New York: Norton.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carmines, E.G., & Zeller, R.A. (1979). *Reliability and validity assessment*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-017. Beverly Hills and London: Sage Publications.

Carroll, J.B. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Cattel, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Centra, J.A. (1973). Effectiveness of student feedback in modifying college intruction. *Journal of Educational Psychology, 65*, 395-401.

Centra, J.A. (1973). *The student instructional report: comparisons with alumni ratings; item reliabilities; the factor structures*. SIR Report No. 3. Princeton, NJ: Educational Testing Service.

Centra, J.A. (1977). Student ratings of instruction and their relationship to student learning. *American Educational Research Journal, 14*, 17-24.

Centra, J.A. (1979). *Determining faculty effectiveness*. San Francisco: Jossey-Bass.

Cohen, P.A. (1980). Effectiveness of student rating feedback for improving college instruction: a meta-analysis of findings. *Research in Higher Education, 13*, 321-341.

Cohen, P.A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309.

Cohen, P.A. (1983). Comment on a selective review of the validity of student ratings of teaching. *Journal of Higher Education, 54*, 448-458.

Conrad, C.F. & Blackburn, R.T. (1985) Program quality in higher education. In: Smart, J.C. (Ed.). *Higher Education: Handbook of Theory and Research*. New York: Agathon Press, Inc.

Cooley, W.W., & Bickel, W.E. (1986). *Decision-Oriented Educational Research*. Boston: Kluwer-Nijhoff Publishing.

Cooley, W.W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis*, 1980, 7-25.

Cooley, W.W., & Lohnes, P.R. (1976). *Evaluation reseach in education*. New York: Irvington Publishers.

Cooper, W.H. (1981). Ubiquitous Halo. *Psychological Bulletin, 90*, 218-244.

Corte, E., de, et al. (1976). *Beknopte Didaxologie*. Groningen: Wolters-Noordhoff.

Costin, F., Greenough, W.T., & Menges, R.J. (1971). Student ratings of college teachers: reliability, validity and usefulness. *Review of Educational Research, 56*, 331-364.

Cousins, J.B. & Leitwood, K.A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*, 331-364.

Cronbach, L.J. (1972). *Essentials of psychological testing*. Third Edition. New York: Harper & Row, Publishers.

Cronbach, L.J., Gleser, G.C., Harinder Nanda., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Daniëls, M. (1985). Over personeelsbeoordeling. *Onderzoek van Onderwijs*.

Darling-Hammond, L., Wise, A.E., & Pease, A.E. (1983). Teacher evaluation in the organizational context: a review of the literature. *Review of Educational Research, 53*, 285-328.

Dewey, J. (1916). *Democracy and Education*. New York: McMillan.

Dowell, D.A., & Neal, J.A. (1982). A selective review of the validity of student ratings of teaching. *Journal of Higher Education, 53*, 51-62.

Doyle, K.O. (1975). *Student evaluation of instruction*. Lexington, MA: D.C. Heath.

Doyle, K.O., & Crichton, L.I. (1978). Student, peer and self-evaluations of college instructors. *Journal of Educational Psychology, 78,* 815-826.

Drenth, P.J.D. (1975). *Inleiding in de testtheorie.* Deventer: Van Loghum Slaterus.

Ebel, R.L. (1951). Estimation of reliability of ratings. *Psychometrika, 16,* 407-424.

Feldman, K.A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher education, 4,* 69-111.

Feldman, K.A. (1977). Consistency and variability among college students in rating their teachers and courses: a review and analysis. *Research in Higher Education, 6,* 223-274.

Fraser, C.E. (1931). *The case method of instruction.* New York: McGraw-Hill.

Frederiksen, C.H., & Rotondo, J.A. (1979). Time-series models and the study of longitudinal change. In: Nesselroade, J.R, & Baltes, P.B. (Ed.). *Longitudinal research in the study of behavior and development.* New York: Academic Press.

Frey, P.W., Leonard, D.W., & Beatty, W.W. (1975). Student ratings of instruction: validation research. *American Educational Research Journal, 12,* 435-447.

Gagne, R.M. (1974). *Essentials of learning and instruction.* Hinsdale, Ill.: Dryden Press.

Gelder, L., van et al. (1975). *Didactische analyse I.* Groningen: Pedagogisch instituut.

Giesbers, J.H.G.I. (1986). De rol van de faculteit bij de kwaliteitsbeheersing van het onderwijs. *Universiteit en Hogeschool, 32,* 284-296.

Gijselaers, W.H. (1983). *Een programma-evaluatie-strategie voor probleemgestuurd onderwijs.* Maastricht: doctoraalscriptie, RL.

Gijselaers, W.H., Schmidt, H.G., & Wijnen, W.H.F.W. (1984). *Bloktoetsgericht studeren.* B.O. 84.26144. Maastricht: Rijksuniversiteit Limburg.

Gijselaers, W.H., & Schmidt, H.G. (1985a). An approach to programme evaluation based on comparative data. In: Goodlad, S. (Ed.). *Accountable autonomy. Perspectives in professional education*. Guildford, Surrey: SRHE.

Gijselaers, W.H., & Schmidt, H.G. (1985b).The development and evaluation of a causal model of problem-based learning. In: Khattab, T., Schmidt, H., Nooman, Z., & Ezzat, E. (Eds.). *Innovation in medical education: an evaluation of its present status*. New York: Springer Publishing Company, in press.

Gilmore, G.M. (1973). *Estimates of reliability coefficients for items and subscales of the Illinois Course Evaluation Questionnaire*. Research Report No. 341. Urbana-Champaign, Il: Measurement and Research Division, Office of Instructional Resources, University of Illinois.

Gilmore, G.M., Kane, M.T., & Naccarato, R.W. (1978). The generalizability of student ratings of instruction: Estimates of teacher and course components. *Journal of Educational Measurement, 15*, 1-13.

Glaser, R. (1976). Components of a psychology of instruction: Towards a science of design. *Review of Educational Research, 46*, 1-24.

Glaser, R. (1984). Education and thinking. The role of knowledge. *American Psychologist, 39*, 93-104.

Goldstein, G., & Hersen, M. (Eds.) (1984). *Handbook of psychological assessment*. New York: Pergamon Press.

Groot, A.D., de, & Naerssen, R.F., van, (1977). *Studietoetsen, construeren afnemen, analyseren*. Deel II. Den Haag: Mouton Publishers.

Guilford, J.P. (1954). *Psychometric Methods*. Second Edition. New York: McGraw-Hill Book Company, Inc.

Guire, K.E., & Kowalski, C.J. (1979). Mathematical description and representation of developmental change functions on the intra- and interindividual levels. In: Nesselroade, J.R, & Baltes, P.B. (Eds.). *Longitudinal research in the study of behavior and development*. New York: Academic Press.

Guthrie, E.R. (1949). The evaluation of teaching. *Educational Record, 30*, 109-115.

Guthrie, E.R. (1954). *The evaluation of teaching: a progress report*. Seattle, WA: University of Washington.

Haertel, G.D., Walberg, H.J., & Weinstein, T. (1983). Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research, 53*, 75-91.

Hamilton, J.D. (1976). The McMaster curriculum: a critique. *British Medical Journal, 1*, 1191-1196.

Harnischfeger, A, & Wiley, D.E. (1976). The teaching learning process in elementary schools: A synoptic view. *Curriculum Inquiry, 6*, 5-43.

Harris, J.W., Horrigan, D.L., Ginther, J.R., & Ham, T.H. (1962). Pilot study in teaching hematology with emphasis on self-education by the students. *Journal of Medical Education, 37*.

Hoger Onderwijs: Autonomie en Kwaliteit. (1985). 's-Gravenhage: Staatsuitgeverij.

Hoger Onderwijs en Onderzoek Plan. Kerndocument. (1987). 's-Gravenhage: Staatsuitgeverij.

Holmes, D.S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of educational Psychology, 63*, 130-133.

Howard, G.S., & Maxwell, S.E. (1980). Correlation between student satisfaction and grades: a case of mistaken causation? *Journal of Educational Psychology, 72*, 810-820.

Howard, G.S., Conway, C.G., & Maxwell, S.E. (1985). Construct Validity of measures of college teaching effectiveness. *Journal of Educational Psychology*, 77, 187-196.

Hoyt, D.P. (1969). *Instructional effectiveness. II. Identifying effective classroom procedures*. Report No. 7. Manhattan, KS: Office of Educational Research, Kansas State University.

Hoyt, D.P. (1973a). Identifying effective educational procedures. *Improving College and University Teaching, 21*, 21-31.

Hoyt, D.P. (1973b). Measurement of instructional effectiveness. *Research in Higher education, 1*, 367-378.

Jones, J. (1981). Students: models of university teaching. *Higher education, 10*, 529-549.

Jöreskog, K.G., & Sörbom, D. (1978). *Lisrel IV: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: International Educational Services.

Kane, M.T., & Brennan, R.L. (1977). The generalizability of class means. *Review of Educational Research,* 77, 267-292.

Katona, G. (1940). *Organizing and memorizing.* New York: Norton.

Kerlinger, F.N. (1973). *Foundations of behavioral research.* New York: Holt, Rinehart, & Winston.

Kim, J., & Mueller, C.W. (1978a). *Introduction to factor analysis. What it is and how to do it.* Sage University Paper series on quantitative applications in the social sciences, series no. 07-013. Beverly Hills and London: Sage Publications.

Kim, J. & Mueller, C.W. (1978b). *Factor analysis. Statistical methods and practical issues.* Sage University Paper series on quantitative applications in the social sciences, series no. 07-014. Beverly Hills and London: Sage Publications.

Kohlan, R.G. (1973). A comparison of faculty evaluations early and late in the course. *Journal of Higher Education, 44,* 587-595.

Kounin, J.S. (1970). *Discipline and group management in classrooms.* New York: Holt, Rinehart & Winston.

Kulik, J.A., & McKeachie, W.J. (1975). The evaluation of teachers in higher education. In F.N. Kerlinger (Ed.), *Review of Research in Education* (Vol. 3, pp. 210-240). Itasca, Il: Peacock.

Kulik, J.A., & Kulik, C.L. (1974). Student ratings of instruction. *Teaching of Psychology, 1,* 51-57.

Kwaliteit van het wetenschappelijk onderwijs. Brenninkmeyer, G. et.al. 's-Gravenhage: Academische Raad (1985).

Larson, J.R. (1979). The limited utility of factor analytic techniques for the study of implicit theories in student ratings of teacher behavior. American Educational Research Journal, 16, 201-211.

Levinson-Rose, J., & Menges, R.J. Improving college teaching: a critical recview of research. *Review of Educational Research, 51,* 403-434.

Levinton, L.C., & Hughes, E.F.X. (1981). Research on the utilization of evaluations. *Evaluation Review, 5,* 525-548.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marsh, H.W. (1977). The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. *American Educational Research Journal, 14,* 441-447.

Marsh, H.W. (1980). Research on students' evaluations of teaching effectiveness: A reply to Vechio. *Instructional Evaluation, 4,* 5-13.

Marsh, H.W. (1982a). SEEQ: a reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52,* 77-95.

Marsh, H.W. (1982b). Validity of students' evaluations of college teaching: a multitrait-multimethod analysis. *Journal of Educational Psychology, 74,* 264-279.

Marsh, H.W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76,* 707-754.

Marsh, H.W., & Cooper, T.L. (1981). Prior subject interest, student's evaluations, and instructional effectiveness. *Multivariate Behavioral Research, 16,* 82-104.

Marsh, H.W., Fleiner, H., & Thomas, C.S. (1975). Validity and usefulness of student evaluations of educational quality. *Journal of Educational Psychology, 74,* 264-279.

Marsh, H.W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. *Journal of Educational Measurement, 20,* 231-248.

Marsh, H.W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal, 21,* 341-366.

Marsh, H.W., & Overall, J.U. (1979). Long-term stability of students' evaluations: A note on Feldman's "Consistency and variability among college students in rating their teachers and courses". *Research in Higher education, 10,* 139-147.

Marsh, H.W., Overall, J.U., & Kesler, S.P. (1979). Validity of student evaluations of instructional effectiveness: a comparison of faculty self evaluations and evaluations by their students. *Journal of Educational Psychology, 71,* 149-160.

Marsh, H.W., & Ware, J.E. (1982). Effects of expressiveness, Content Coverage, and incentive on multidimensional student rating scales: new interpretations of the Dr. Fox effect. *Journal of Educational Psychology, 74,* 126-134.

Maslow, A.H., & Zimmerman, W. (1956). College teaching ability, scholarly activity, and personality. *Journal of Educational Psychology, 47,* 185-189.

McKeachie, W.J. (1979). Student ratings of faculty: a reprise. *Academe, 65,* 384-397.

McKeachie, W.J., Lin, Y-G., & Mann, W. (1971). Student ratings of teacher effectiveness: validity studies. *American Educational Research Journal, 8,* 435-445.

Meerling. (1981). *Methoden en technieken van psychologisch onderzoek. Deel 1: Model, observatie en beslissing.* Meppel: Boom.

Mitchell, D.E., & Kerchner, C.T. (1983). Collective bargaining and teacher policy. In: Shulman, L.S., & Sykes, G. (Eds.), *Handbook of teaching and policy.* New York: Longman.

Moust, J.H.C., Grave, W.S., de, & Gijselaers, W.H. (1985). The tutor role a neglected variable in the implementation of problem-based learning. In: Khattab, T., Schmidt, H., Nooman, Z., & Ezzat, E. (Eds.). *Innovation in medical education: an evaluation of its present status.* New York: Springer Publishing Company, in press.

Moust, J.H.C., & Schmidt, H.G. (1985). Preparing faculty members and students for problem-based learning. In: Schmidt, H.G., Vries, M., de, & Lipkin, M. (Eds.). *Education of tomorrow's medicine today.* New York: Springer Publishing Company.

Murray, H.G. (1975). Predicting student ratings of college teaching from peer ratings of personality traits. *Teaching of Psychology, 2,* 66-69.

Naftulin, D.H., Ware, J.E., & Donnely, F.A. (1973). The docter Fox lecture: a paradigm of educational seduction. *Journal of Medical Education, 48,* 630-635.

Neame, R.L.B. (1984). Problem-centred learning in medical education: the role of the context in the development of process skills. In: Schmidt, H.G., De Volder, M.L. (Eds.). *Tutorials in problem-based learning*. Assen: Van Gorcum.

Neufeld, V.R., & Barrows, H.S. (1974). The McMaster philosophy': an approach to medical education. *Journal of Medical Education, 49*, 1040-1050.

Neve, M.F., de, & Janssen, P.J. (1982). Validity of student evaluation of instruction. *Higher Education, 11*, 543-552.

Nie, N.H., et al. (1983). *SPSSX*. New York: McGraw-Hill.

Nisbett, R.E., & Wilson, T.D. (1977). The halo effect: evidence for unconscious alteration of judgements. *Journal of Personality and Social Psychology, 35*, 250-256.

Ory, J.C., & Braskamp, L.A. (1981). Faculty perceptions of the quality and usefulness of three types of evaluative information. *Research in Higher Education, 15*, 271-281.

Os, W. van. (1987). *Evaluatie in het hoger onderwijs: controle op en verbetering van de kwaliteit van hoger onderwijs*. Groningen: Wolters-Noordhoff (Hoger Onderwijs Reeks).

Page, C.F. (1974). *Student evaluation of teaching: the American experience*. London: Society for Research into Higher Education.

Pambookian, H.S. (1974). The initial level of student evaluation of instruction as a source of influence on instructor change after feedback. *Journal of Educational Psychology, 71*, 856-865.

Parlett, M., & Hamilton, D. (1977). Evaluation as illumination: a new approach to the study of innovatory programs. In: Parlett, M., & Dearden G. *Introduction to illuminative evaluation: studies in higher education*.

Posthumus, K. (1968). *De universiteit -doelstellingen, functies en structuren*. Discussienota Posthumus, studenteneditie. 's-Gravenhage.

Remmers, H.H., & Weisbrodt, J.A. (1964). *Manual of instructions for Purdue rating scale of instruction*. Purdue, IN: Purdue Research Foundation.

Rodin, M., & Rodin, B. (1972). Student evaluation of teachers. *Science, 177*, 1164-1166.

Rogosa, D., Floden, R., & Willet, J.B. (1984). Assessing the

stability of teacher behavior. *Journal of Educational Psychology, 76,* 1000-1027.

Rooijen, L., van, & Vlaander, G.P.J. (1983). Het optreden en uitblijven van halo-effecten in de oordelen van studenten over docenten. *Tijdschrift voor Onderwijsresearch, 8,* 157-171.

Rotem, A., & Glasman, N.S. (1979). On the effectiveness of students' evaluative feedback to university instructors. *Review of Educational Research, 49,* 497-511.

Rudduck, J. (1978). *Learning through small group discussion.* Research into higher education monographs. Guildford, Surrey: SRHE.

Russel, B. (1946). *Geschiedenis Der Westerse Filosofie in samenhang met politieke en sociale omstandigheden van de oudste tijden tot heden.* Katwijk aan Zee: Servire.

Saal, F.E., Downey, R.G., & Lahey, M.A. (1980). Rating the ratings: assessing the psychometric quality of rating data. *Psychological Bulletin, 88,* 413-428.

Schmidt, H.G. (1979). Leren met problemen, een inleiding in probleemgestuurd onderwijs. In: Vroon, A.G., et al. *Handboek voor de onderwijspraktijk 1.* Deventer: Van Loghum Slaterus.

Schmidt, H.G. (1981). *Vragenlijst voor de beoordeling van het onderwijs in de geneeskunde.* Maastricht: RL.

Schmidt, H.G. (1982). Activatie en herstructurering van voorkennis en hun effect op de verwerking van tekst: gegevens uit recall. In: H.G. Schmidt *Activatie van voorkennis, intrinsieke motivatie en de verwerking van tekst.* Academisch proefschrift, Maastricht.

Schmidt, H.G., & Bouhuijs, P.A.J. (1980) *Onderwijs in taakgerichte groepen.* Utrecht: Spectrum.

Schmidt, H.G., & Volder, M.L., de, (1984). *Tutorials in problem-based learning.* A new direction in teaching the health professions. Assen: Van Gorcum.

Schmidt, H.G., De Volder, M.L., Gijselaers, W.H., & Kerkhofs, L.M.M. (1984). Een positief verband tussen studiejaar en tentamenresultaat, en de rol van toenemende voorkennis. *Tijdschrift voor Onderwijsresearch, 9,* 183-188.

Schröer, C.A.P. (1985). *De organisatie van blokplanning-groepen.* B.O. 85-6029. Maastricht: Rijksuniversiteit Limburg.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability theory of measures of teacher behavior. *Review of Educational Research, 46*, 553-611.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Snellen-Balendong, H.A.M. (1985). Development of formats for tasks and problems in problem-based learning. In: Grave, W.S., de, Moust, J.H.C., & Schmidt, H.G. *Tutorials in problem-based learning*. Volume II. Maastricht, Rijksuniversiteit Limburg.

Sweeney, G.D., & Mitchell, D.L.M. (1975). An introduction to the study of medicine: phase I of the McMaster M.D. program. *Journal of Medical Education, 50*, 70-77.

Tabachnik, B.G., & Fidell, L.S. (1983). *Using multivariate statistics*. New York: Harper & Row Publishers.

Tinsley, H.E.A., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22*, 358-376.

Tyler, R.W. (1934). *Constructing achievement tests*. Colombus, Ohio: Ohio State University.

Tyler, R.W. (1950). *Basic principles of curriculum and instruction*. Chicago, Ill.: University of Chicago Press.

Verwijnen, G.M., et al. (1982). The evaluation system at the medical school of Maastricht. *Assessment and Evaluation in Higher Education, 3*, 225-244.

Volder, M.L. de. (1981). Functioning of discussion groups and their tutors: structural relations with tutor characteristics and academic achievement of groups. *Onderzoek van Onderwijs*, 8, Rijksuniversiteit Limburg, Maastricht.

Vollmer, F. (1986). The relationship between expectancy and academic achievemnt. How can it be explained? *British Journal of Educational Psychology, 56*, 64-74.

Ware, J.E., & Williams, R.G. (1977). An extended visit with Dr. Fox: Validity of student ratings of instruction after repeated exposure to a lecturer. *American Educational Research Journal, 14*, 449-457.

Wilson, R.C. (1986). Improving faculty teaching. Effective use of student evaluations and Consultants. *Journal of Higher Education, 59*, 196-211.

Winer, B.J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

CURRICULUM VITAE.

De schrijver van dit proefschrift werd op 7 juli 1959 geboren te Heerlen. Hij voltooide zijn Atheneum-b opleiding in 1977 aan het St. Antonius Doctor College te Kerkrade. In datzelfde jaar begon hij met zijn studie Interdisciplinaire Onderwijskunde aan de Katholieke Universiteit te Nijmegen. Deze studie werd in 1983 voltooid met als specialisatie methoden van onderwijskundig onderzoek en als bijvakken leerstoornissen en economische sociologie. Sedert 1982 is hij als universitair docent verbonden aan de vakgroep Onderwijsontwikkeling en Onderwijsresearch van de Rijksuniversiteit Limburg.